Uniform Distribution Theory,
Vol. 10 (2015), no. 2

(photo: *P. Liardet*)

Professor Niederreiter giving the lecture
on the International Conference on Uniform Distribution Theory,
Marseilles, CIRM, January 21–25, 2008.

Uniform Distribution Theory
Vol. 10 (2015), no. 2

Dedicated to
Professor Harald Niederreiter
on the occasion of his seventieth birthday

CONTENTS

Articles

Vol. 10, no. 2 **2015**

Uniform Distribution Theory **10** (2015), no.2, 1–10

uniform distribution theory

STATISTICAL DISTRIBUTION OF ROOTS MODULO PRIMES OF AN IRREDUCIBLE AND DECOMPOSABLE POLYNOMIAL OF DEGREE 4

Yoshiyuki Kitaoka

Dedicated to Professor Harald Niederreiter on the occasion of his 70th birthday

ABSTRACT. For an irreducible polynomial $f(x) = (x^2 + ax)^2 + b(x^2 + ax) + c$ of degree 4 and a natural number L, we propose a conjecture of distribution of roots r_1, r_2, r_3, r_4 of f modulo a prime p satisfying $r_i \equiv 0 \bmod L$ and $0 \le r_i \le pL - 1$.

Communicated by Shigeki Akiyama

1. Introduction

Let

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{Z}[x]$$

be an irreducible monic polynomial with integer coefficients, and let L be a natural number. Put

$$Spl(f) = \{p \mid f(x) \bmod p \text{ is completely decomposable}\},$$

where a letter p denotes prime numbers larger than L. This is an infinite set, and the natural density is given by Chebotarev's density theorem. For a prime $p \in Spl(f)$, we can take n integers $r_1, \ldots, r_n \in \mathbb{Z}$ such that

$$\begin{cases} f(r_i) \equiv 0 \bmod p, \\ \quad r_i \equiv 0 \bmod L, \qquad (i = 1, \ldots, n) \\ 0 \le r_i \le pL - 1, \end{cases} \tag{1}$$

2010 Mathematics Subject Classification: 11K.
Keywords: polynomial, roots modulo prime, distribution.

by Chinese Remainder Theorem. Then we have $a_{n-1} + \sum r_i \equiv 0 \bmod p$, hence there exists an integer $C_p(f)$ such that

$$a_{n-1} + \sum_{i=1}^{n} r_i = C_p(f)p\,. \tag{2}$$

We note that $-a_{n-1}$ is the global trace of a root α of $f(x) = 0$ and $\sum r_i$ is the sum of local traces in $\mathbb{Q}(\alpha) \otimes \mathbb{Q}_p$ modulo p, hence $C_p(f)$ involves the difference of the global trace and the sum of local traces. The condition

$$1 \leq C_p(f) < nL \tag{3}$$

holds with finitely many exceptional primes p, and we studied the distribution of $C_p(f)$ for an irreducible and indecomposable[1] polynomial f in the previous paper [5]. Putting

$$Pr_X(f, L)[k] = \frac{\#\{p \in Spl_X(f) \mid C_p(f) = k\}}{\#Spl_X(f)},$$

where $Spl_X(f) = \{p \in Spl(f) \mid p \leq X\}$, we are concerned with the limit

$$Pr(f, L)[k] := \lim_{X \to \infty} Pr_X(f, L)[k]\,.$$

Numerical data suggest that the limit exists, and we gave several observations ([1] in the case related to the decimal expansion of a rational number, [2], [3], [4] in the case of $L = 1$, and [5] in the case of $L > 1$ and an irreducible and indecomposable polynomial). By a remark on $C_p(f)$ above, $Pr(f, L)[k] = 0$ if $k \leq 0$ or $k \geq nL$.

In the subsection 2.2.1 of [5], we gave conjectures for an irreducible and decomposable(= reduced there) polynomial of degree 4, but they were observations based on insufficient data. In fact, one of them is false. Here we correct it and give a definite result in the easiest case and conjectures based on more data in the next section. In the third section, we give a proof of the easiest case and theoretical partial evidences of conjectures.

2. Conjectures

First, let us recall some necessary results in [5]. Let

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

be an irreducible polynomial with integer coefficients. Hereafter, $\mathbb{Q}(f)$ denotes

[1] A polynomial f is called indecomposable unless $f(x) = g(h(x))$ holds for polynomials g, h with $1 < \deg g < \deg f$. In [5], it was called non-reduced.

the Galois extension of the rational number field $\mathbb{Q}$ generated by all roots of $f(x) = 0$ and ζ_T is a primitive Tth root of unity.

The following is Proposition 1 in [5]:

Proposionion 1. *Let f be a polynomial above, and let L, j be natural numbers, and put $N = (a_{n-1}, L)$ and $T = L/N$. We denote Euler's function by φ. If $T = 1$, i. e., $a_{n-1} \equiv 0 \bmod L$, then we have*

$$\lim_{X\to\infty} \sum_{k\equiv j \bmod L} Pr_X(f, L)[k] = \begin{cases} 1 & \text{if } j \equiv 0 \bmod L, \\ 0 & \text{otherwise.} \end{cases}$$

If $T > 1$, then we have

$$\lim_{X\to\infty} \sum_{k\equiv j \bmod L} Pr_X(f, L)[k] = \frac{[\mathbb{Q}(f) \cap \mathbb{Q}(\zeta_T) : \mathbb{Q}]}{\varphi(T)} \text{ or } 0,$$

where the limit is not zero if and only if (i) $(j, L) = N$ *and* (ii) *mappings $\zeta_T \to \zeta_T^{a_{n-1}/N}$ and $\zeta_T \to \zeta_T^{j/N}$ induce the same automorphism on the subfield $\mathbb{Q}(f) \cap \mathbb{Q}(\zeta_T)$ of $\mathbb{Q}(\zeta_T)$.*

Although the existence of the limit of each factor $Pr_X(f, L)[k]$ is a conjecture, the limit of the sum $\sum_{k\equiv j \bmod L} Pr_X(f, L)[k]$ exists. Hereafter as in the proposition, we put

$$N := (a_{n-1}, L),\ T := L/N, \tag{4}$$

hence $(a_{n-1}/N, T) = 1$. The proposition says that the non-vanishing condition $Pr(f, L)[k] \neq 0$ implies $(k, L) = N$, hence we introduce the shrunk density

$$SPr(f, L)[k] := Pr(f, L)[kN] \quad (1 \le k < nT).$$

Under the basic conjecture of the existence of the limit $Pr(f, L)$, the condition $SPr[k] \neq 0$ implies that (i) $1 \le k < nT$ and (ii) $(k, T) = 1$, and (iii) k and a_{n-1}/N induce the same automorphism on the field $\mathbb{Q}(f) \cap \mathbb{Q}(\zeta_T)$, and furthermore

$$\sum_{j\equiv k \bmod T} SPr[j] = \frac{[\mathbb{Q}(f) \cap \mathbb{Q}(\zeta_T) : \mathbb{Q}]}{\varphi(T)} \tag{5}$$

for any integer k satisfying the three conditions above.

From now on, we specialize a polynomial f to *an irreducible and decomposable polynomial of degree* 4, i. e.,

$$f(x) = (x^2 + ax)^2 + b(x^2 + ax) + c \quad (a, b, c \in \mathbb{Z}), \tag{6}$$

whence $n = 4, a_{n-1} = 2a$ and $(2a/N, T) = 1$.

The case of $T = 1$, i. e., $L \mid 2a$ is as follows:

Theorem 1. *If $a \equiv 0 \bmod L$, then*

$$SPr(f, L)[2] = 1, \quad SPr(f, L)[1] = SPr(f, L)[3] = 0.$$

Conjecture 1. *If $2a \equiv 0 \bmod L, a \not\equiv 0 \bmod L$, then*

$$SPr(f, L)[2] = 1/2, \quad SPr(f, L)[1] = SPr(f, L)[3] = 1/4.$$

The proof and the comment are given in the next section.

Next, suppose that $T > 1$ and put

$$f_2(x) = -2x^2 + 8Tx - 6T^2, \quad f_3(x) = f_2(x) + x^2.$$

Fixing T, we introduce basic vectors $v_i[k]$ for $k = 1, \ldots, T-1$:

$$\begin{aligned}
v_1[k] &:= (0, 0, 0, 0),\\
v_2[k] &:= \big(k^2, f_2(T+k), (2T-k)^2, 0\big),\\
v_3[k] &:= \big(0, (T+k)^2, f_2(2T-k), (T-k)^2\big),\\
v_4[k] &:= \big(k^2, f_3(T+k), f_3(2T-k), (T-k)^2\big).
\end{aligned}$$

We expect that for $1 \le k < T$, a vector

$$V[k] := \big(SPr[k], SPr[T+k], SPr[2T+k], SPr[3T+k]\big)$$

is proportional to one of vectors $v_i[k]$. Since the sum of entries of the second, the third and the last is equal to $4T^2$, $4T^2$ and $8T^2$, respectively, the equation (5) suggests that $V[k]$ is equal to one of

$$v_1[k], \quad 2q_T v_2[k], \quad 2q_T v_3[k], \quad q_T v_4[k],$$

where

$$q_T := \frac{[\mathbb{Q}(f) \cap \mathbb{Q}(\zeta_T) : \mathbb{Q}]}{8T^2\varphi(T)}.$$

The data suggest that the proportional constant is independent of k, hence v_4 does not appear together with v_2, v_3 at the same time. We note that $v_2[k] + v_3[k] = v_4[k]$ and entries in $v_i[k]$ are positive if "0" is not put.

Let F be an abelian extension of $\mathbb{Q}$ and let C_F be the conductor of F, that is the least positive integer C_F such that $\mathbb{Q}(\zeta_{C_F})$ contains F. If an integer k is relatively prime to C_F, we denote by $[[k]]$ an automorphism of F induced by $\zeta_{C_F} \to \zeta_{C_F}^k$.

Now we can state conjectures in case of $T > 1$. Note that the order of the Galois group $Gal(\mathbb{Q}(f)/\mathbb{Q})$ is $4, 8$ and Proposition 1 implies that $V[k] \neq (0,0,0,0)$ if and only if $(k, T) = 1$ and $[[k]] = [[2a/N]]$ on $\mathbb{Q}(f) \cap \mathbb{Q}(\zeta_T)$.

Hereafter integers k are supposed to satisfy

$$1 \le k \le T-1,\ (k, T) = 1 \quad \textit{and} \quad [[k]] = [[2a/N]] \textit{ on } \mathbb{Q}(f) \cap \mathbb{Q}(\zeta_T). \tag{7}$$

If one of the above is not satisfied, we have $V[k] = (0, 0, 0, 0)$ by Proposition 1.

(I) The case of $T \equiv 1 \bmod 2$ [2]:
$V[k] = q_T v_4[k]$ does not occur, and

$$V[k] = 2q_T \times \begin{cases} v_2[k] & \text{if } k \equiv 2a/N \bmod 2, \\ v_3[k] & \text{if } k \not\equiv 2a/N \bmod 2. \end{cases}$$

(II) The case of $T \equiv 0 \bmod 2$:
Let F be the maximal abelian subfield of $\mathbb{Q}(f)$, which is quartic.

(II.a) The case of $[F \cap \mathbb{Q}(\zeta_{2T}) : F \cap \mathbb{Q}(\zeta_T)] = 2$:

$$V[k] = 2q_T \times \begin{cases} v_2[k] & \text{if } [[k]] = [[2a/N]] \text{ on } F \cap \mathbb{Q}(\zeta_{2T}), \\ v_3[k] & \text{if } [[k]] \neq [[2a/N]] \text{ on } F \cap \mathbb{Q}(\zeta_{2T}). \end{cases}$$

(II.b) The case of $[F \cap \mathbb{Q}(\zeta_{2T}) : F \cap \mathbb{Q}(\zeta_T)] \neq 2$:
$V[k] = q_T v_4[k]$ holds for every k.

The (conjectural) density depends not on the field $\mathbb{Q}(f)$ but on the maximal abelian subfield F.

Our checking method by pari/gp[3] is to watch that numerical data $Pr_X(f, L)$ multiplied by $8T^2\varphi(T)$ approach the conjectural densities multiplied by the same $8T^2\varphi(T)$, which are integers. Let us give numerical data of ratios $Pr_X(f, L)[k]$ to the conjectural density $Pr[k]$ in case of $a = 4$, $b = 13$, $c = 41$, $N = 8$, $21 \leq T \leq 30$, $L = NT$, $X = 10^{10}$: By $\mathbb{Q}(f) = \mathbb{Q}(\zeta_5)$, the maximal abelian subfield F is $\mathbb{Q}(f)$ itself and $F \cap \mathbb{Q}(\zeta_T) = \mathbb{Q}(\zeta_{(5,T)})$. Hence, in the case of $5 \mid T$, the condition $[[k]] = [[2a/N]]$ is $k \equiv 2a/N \bmod 5$. We define er by

$$er = \max_{\substack{1 \leq k \leq 4L-1 \\ Pr[k] \neq 0}} |Pr_X(f, L)[k]/Pr[k] - 1|.$$

In the following table, T is on the upper row, $1/q_T$ is on the middle line and er is on the lower row, where the value er is rounded off to the third decimal place.

T	21	22	23	24	25
$1/q_T$	42336	38720	93104	36864	25000
er	0.011	0.048	0.022	0.009	0.004

T	26	27	28	29	30
$1/q_T$	64896	104976	75264	188384	14400
er	0.027	0.034	0.017	0.026	0.003

[2] In [5], the case of $2a/N \equiv 1 \bmod 2$ was missing.

[3] The PARI Group, PARI/GP version 2.7.0 Bordeaux, 2014, `http://pari.math.u-bordeaux.fr/`

We note that in case of $T \equiv 0 \bmod 2$,

$$\begin{aligned}&[F \cap \mathbb{Q}(\zeta_{2T}) : F \cap \mathbb{Q}(\zeta_T)] = 2\\ \Leftrightarrow &\begin{cases} 2 \,|\, C_F,\ F \cap \mathbb{Q}(\zeta_{2T}) \neq \mathbb{Q} & \text{if } [F \cap \mathbb{Q}(\zeta_T) : \mathbb{Q}] = 1,\\ C_F = 2(C_F, T) & \text{if } [F \cap \mathbb{Q}(\zeta_T) : \mathbb{Q}] = 2. \end{cases}\end{aligned}$$

P r o o f. Suppose first, $[F \cap \mathbb{Q}(\zeta_{2T}) : F \cap \mathbb{Q}(\zeta_T)] = 2$. If $F \cap \mathbb{Q}(\zeta_T) = \mathbb{Q}$, then two equations

$$\begin{aligned}[F(\zeta_{2T}) : \mathbb{Q}] &= [F : F \cap \mathbb{Q}(\zeta_{2T})][\mathbb{Q}(\zeta_{2T}) : F \cap \mathbb{Q}(\zeta_{2T})] \cdot 2\\ &= 2 \cdot \varphi(2T)/2 \cdot 2 = \varphi(4T)\end{aligned}$$

and

$$\begin{aligned}[F(\zeta_{2T}) : \mathbb{Q}] &= [F(\zeta_{2T}) : F(\zeta_T)][F : \mathbb{Q}][\mathbb{Q}(\zeta_T) : \mathbb{Q}]\\ &= [F(\zeta_{2T}) : F(\zeta_T)] \cdot 4\varphi(T) = [F(\zeta_{2T}) : F(\zeta_T)]\varphi(4T)\end{aligned}$$

imply

$$[F(\zeta_{2T}) : F(\zeta_T)] = 1, \quad \text{i. e.,}\ \ F(\zeta_{2T}) = F(\zeta_T),$$

hence

$$\mathbb{Q}(\zeta_{C_F}, \zeta_{2T}) = \mathbb{Q}(\zeta_{C_F}, \zeta_T).$$

Thus C_F should be even, since T is even. If $F \cap \mathbb{Q}(\zeta_T) \neq \mathbb{Q}$, then $F \cap \mathbb{Q}(\zeta_T)$ is quadratic and $F \cap \mathbb{Q}(\zeta_{2T})$ is quartic, that is $F \cap \mathbb{Q}(\zeta_{2T}) = F$, hence

$$F \not\subset \mathbb{Q}(\zeta_T) \quad \text{and} \quad F \subset \mathbb{Q}(\zeta_{2T}),$$

which imply

$$C_F = 2(C_F, T).$$

Conversely, suppose that $2 \,|\, C_F$, $F \cap \mathbb{Q}(\zeta_{2T}) \neq \mathbb{Q}, [F \cap \mathbb{Q}(\zeta_T) : \mathbb{Q}] = 1$; we have only to show $F \not\subset \mathbb{Q}(\zeta_{2T})$. If $F \subset \mathbb{Q}(\zeta_{2T})$, then we have

$$[F(\zeta_T) : \mathbb{Q}] = [F : \mathbb{Q}][\mathbb{Q}(\zeta_T) : \mathbb{Q}] = 4\varphi(T),$$

hence

$$\varphi(2T) = [\mathbb{Q}(\zeta_{2T}) : \mathbb{Q}] = [F(\zeta_{2T}) : \mathbb{Q}] = [F(\zeta_{2T}) : F(\zeta_T))] \cdot 4\varphi(T),$$

which is a contradiction. Last, assume $C_F = 2(C_F, T), [F \cap \mathbb{Q}(\zeta_T) : \mathbb{Q}] = 2$; the odd part of C_F divides T and the 2-factor of C_F is twice the 2-factor of T, i. e., $C_F \,|\, 2T$. Thus $F \subset \mathbb{Q}(\zeta_{C_F}) \subset \mathbb{Q}(\zeta_{2T})$, that is $F \cap \mathbb{Q}(\zeta_{2T}) = F$. □

3. Proof of Theorem 1 and comments

We recall that an irreducible polynomial f is given by (6) and note that the equation $x^2+ax=(-x-a)^2+a(-x-a)$ implies the equivalence of

$$f(r)\equiv 0 \bmod p \quad \text{and} \quad f(-r-a)\equiv 0 \bmod p.$$

A basic lemma is

Lemma 1. *Let an integer r be a root of $f(x)\equiv 0 \bmod p$ with (1); then we can take an integer δ so that for $R:=p\delta-r-a$*

$$\begin{gathered} f(R)\equiv 0 \bmod p,\ 0\le R<pL,\ R\equiv 0 \bmod L, \\ R\not\equiv r \bmod p, \\ 0<\delta<2L, \end{gathered} \tag{8}$$

with finitely many exceptional primes p.

P r o o f. The existence of an integer δ satisfying (8) for a prime $p\,(>L)$ follows from Chinese Remainder Theorem. Suppose $R\equiv r \bmod p$ for an odd prime p; then we have $2r\equiv -a \bmod p$, therefore for an irreducible polynomial $F(x):=2^4f(x/2)$ with integer coefficients, we have $F(-a)\equiv F(2r)\equiv 2^4f(r)\equiv 0 \bmod p$. Thus, such a prime p is a divisor of non-zero integer $F(-a)$, hence the possibility of p is finite. Next, the condition $0\le R<pL$ implies $(\delta-L)p-a<r\le p\delta-a$. By the assumption $0\le r<pL$, we have $(\delta-2L)p<a\le p\delta$, which implies $a/p\le\delta<a/p+2L$, i. e., $0\le\delta\le 2L$ with finitely many exceptional primes p. Suppose $\delta=0$ for infinitely many primes p; then $0\le R=-r-a$, i. e., $0\le r\le -a$ follows for infinitely many primes. Thus there is an integer $M\,(=r)$ between 0 and $-a$ such that $f(M)\equiv 0 \bmod p$ for infinitely many primes, which implies a contradiction $f(M)=0$. Last, suppose $\delta=2L$ for infinitely many primes p; then $R=2Lp-r-a<pL$, i. e., $-a\le r-pL<0$ follows for infinitely many primes. Thus there is an integer $M\,(=r-pL)$ such that $f(M)\equiv f(r)\equiv 0 \bmod p$ for infinitely many primes, which implies a contradiction $f(M)=0$. □

Proposionion 2. *For a prime $p\in Spl(f)$, let $r_1,\dots,r_4$ be roots of $f(x)\equiv 0 \bmod p$ with (1), and using the previous lemma, we may suppose that they satisfy*

$$r_1+r_3=\delta_1p-a,\ r_2+r_4=\delta_2p-a\,(0<\delta_1\le\delta_2<2L). \tag{9}$$

Then we have $\delta_1=\delta_2=C_p(f)/2$ or $\delta_1=(C_p(f)-L)/2$, $\delta_2=(C_p(f)+L)/2$, where $C_p(f)$ is defined by (2).

P r o o f. Since the assumption $r_i \equiv 0 \bmod L$ $(i = 1, \dots, 4)$ implies $\delta_1 p - a \equiv \delta_2 p - a \equiv 0 \bmod L$, we have $\delta_1 \equiv \delta_2 \bmod L$, assuming $p > L$. Thus $\delta_1 = \delta_2$ or $\delta_2 = \delta_1 + L$ follows from $0 < \delta_1 \leq \delta_2 < 2L$, and by $(\delta_1 + \delta_2)p = 2a + \sum r_i = C_p(f)p$, we get the statement of the proposition. □

(1) and (2) imply a fundamental relation $C_p(f)p \equiv 2a \bmod L$, hence for some natural number k

$$C_p(f) = kN, \quad \text{and} \quad (k, T) = 1,\ kp \equiv 2a/N \bmod T. \tag{10}$$

Moreover, we have

Corollary 1. *Continuing the previous proposition, we have, for $C_p(f) = kN$ above*

$$\begin{cases} kp \equiv 2a/N \bmod 2T & \text{if} \quad \delta_2 = \delta_1, \\ (k-T)p \equiv 2a/N \bmod 2T & \text{if} \quad \delta_2 = \delta_1 + L. \end{cases}$$

If either $C_p(f) \leq L$ or $C_p(f) \geq 3L$, then only the case $\delta_2 = \delta_1$ holds.

P r o o f. By (1) and (9), we have $\delta_1 p \equiv a \bmod L$, hence

$$\begin{aligned} kN/2 \cdot p \equiv a \bmod L \quad &\text{if} \quad \delta_2 = \delta_1, \\ (kN - L)/2 \cdot p \equiv a \bmod L \quad &\text{if} \quad \delta_2 = \delta_1 + L. \end{aligned}$$

Since $L = NT$, and $2a$ is divisible by N, the statement follows easily. If the condition $\delta_2 = \delta_1 + L$ happens, then the previous proposition implies

$$(C_p(f) - L)/2 > 0 \quad \text{and} \quad (C_p(f) + L)/2 < 2L, \quad \text{i. e.,} \quad L < C_p(f) < 3L,$$

which completes the proof. □

The corollary says that if either $C_p(f) \leq L$ or $C_p(f) \geq 3L$, we have a stronger condition $kN/2 \cdot p \equiv a \bmod L$ than (10), which elucidates the entry "0" in vectors v_2, v_3 as stated later.

P r o o f o f T h o e r e m 1. By the assumption $a \equiv 0 \bmod L$, δ_1, δ_2 in (9) are divisible by L, hence are equal to L. Thus we have $C_p(f) = 2L$, hence

$$SPr(f, L)[2] = Pr(f, L)[2N] = Pr(f, L)[2L] = 1$$

and

$$SPr(f, L)[1] = SPr(f, L)[3] = 0.$$

□

Next, let us study the meaning of Conjecture 1, i. e.,

$$SPr(f, L) = (1/4, 1/2, 1/4) \quad \text{if} \ \ 2a \equiv 0 \bmod L \quad \text{but} \ \ a \not\equiv 0 \bmod L\,.$$

We use notations in Proposition 2. By the equation (2), we see $C_p(f) \equiv 0 \bmod L$, which implies $C_p(f) = L, 2L$ or $3L$ by (3). We see that $\delta_1 = \delta_2 = L/2$ holds in

the case of $C_p(f) = L$, second $\delta_1 = L/2, \delta_2 = 3L/2$ in the case of $C_p(f) = 2L$ since $\delta_1 = \delta_2 = L$ with (9) implies a contradiction $a \equiv 0 \bmod L$, and last $\delta_1 = \delta_2 = 3L/2$ holds in the case of $C_p(f) = 3L$. Thus we have

$$2a + \sum r_i = Lp \times \begin{cases} 1 & \text{if} \quad \delta_1 = \delta_2 = L/2, \\ 2 & \text{if} \quad \delta_1 = L/2, \delta_2 = 3L/2, \\ 3 & \text{if} \quad \delta_1 = \delta_2 = 3L/2. \end{cases}$$

Here let us see that $\delta_2 = L/2$ (resp. $\delta_2 = 3L/2$) is equivalent to $r_2, r_4 \in [0, Lp/2]$ (resp. $r_2, r_4 \in [Lp/2, Lp]$) except finitely many primes p. Suppose $\delta_2 = L/2$; then $r_2, r_4 \leq r_2 + r_4 = Lp/2 - a$ implies $r_2, r_4 \in [0, Lp/2 - a]$. By using the pigeon hole principle as in the proof of Lemma 1, we see $r_2, r_4 \in [0, Lp/2]$ except finitely many primes p. The equivalence of the conditions $\delta_1 = L/2$ and $r_1, r_3 \in [0, Lp/2]$ is similar. If $\delta_2 = 3L/2$, then the condition $r_2 < pL$ implies

$$r_4 = 3Lp/2 - a - r_2 > Lp/2 - a \quad \text{and similarly} \quad r_2 > Lp/2 - a,$$

hence we have $r_2, r_4 \in [Lp/2, Lp]$ except finitely many primes p, similarly. Thus, if pairs r_1, r_3 and r_2, r_4 distribute uniformly on $[0, Lp/2]$ and $[Lp/2, Lp]$, we have $SPr(f, L) = (1/4, 1/2, 1/4)$.

Let us assume $T > 1$ hereafter, and we show that cases of the density zero in the conjecture are affirmative.

We are still assuming $1 \leq k \leq T - 1$.
The case of $T \equiv 1 \bmod 2$: We have to show

$$SPr[3T + k] = 0 \qquad \text{if } \ k \equiv 2a/N \bmod 2, \tag{11}$$

$$SPr[k] = 0 \qquad \text{if } \ k \not\equiv 2a/N \bmod 2. \tag{12}$$

By Corollary 1, we have $\delta_2 = \delta_1$ for $C_p(f) = kN, (3T + k)N$. Thus the condition $SPr[3T + k] = Pr(f, L)[(3T + k)N] \neq 0$ implies $(3T + k)p \equiv 2a/N \bmod 2T$, which implies $1 + k \equiv 2a/N \bmod 2$. This concludes (11).
Suppose that $SPr[k] = Pr(f, L)[kN] \neq 0$; then we have $kp \equiv 2a/N \bmod 2T$, which implies $k \equiv 2a/N \bmod 2$, and (12) has been proved.

Let us assume that $T \equiv 0 \bmod 2$. Keeping notations in the previous section, we must show that in case of $[F \cap \mathbb{Q}(\zeta_{2T}) : F \cap \mathbb{Q}(\zeta_T)] = 2$,

$$SPr[3T + k] = 0 \qquad \text{if } \ [[k]] = [[2a/N]] \text{ on } F \cap \mathbb{Q}(\zeta_{2T}), \tag{13}$$

$$SPr[k] = 0 \qquad \text{if } \ [[k]] \neq [[2a/N]] \text{ on } F \cap \mathbb{Q}(\zeta_{2T}). \tag{14}$$

We are assuming that an integer k is relatively prime to T and T is even, thus $[[k]]$ is well-defined on $F \cap \mathbb{Q}(\zeta_{2T})\, (\subset \mathbb{Q}(\zeta_{2T}))$.

On (13): Suppose $SPr[3T + k] = Pr[(3T + k)N] \neq 0$; we have $\delta_2 = \delta_1$, hence $(T + k)p \equiv 2a/N \bmod 2T$ for a prime p with $C_p(f) = (3T + k)N$,

hence $\zeta_{2T}^{2a/N} = -\zeta_{2T}^{kp} = -\sigma(\zeta_{2T})^k$, where σ is a Frobenius automorphism of p at $\mathbb{Q}(f)(\zeta_{2T})$. The condition $p \in Spl(f)$ implies that σ is the identity mapping on $F\,(\subset \mathbb{Q}(f))$. Since $K := F \cap \mathbb{Q}(\zeta_{2T}) \neq F \cap \mathbb{Q}(\zeta_T)$, there is an element $\alpha \in K$, which is not expressed by a linear combination of powers ζ_T with rational coefficients, therefore $\alpha^{[[2a/N]]} \neq \sigma(\alpha)^{[[k]]} = \alpha^{[[k]]}$, that is $[[2a/N]] \neq [[k]]$ on K.

On (14): Suppose $SPr[k] = Pr[kN] \neq 0$; then we have $kp \equiv 2a/N \bmod 2T$. Thus we have $[[k]] = [[2a/N]]$ by the fact that p decompose at $F(\subset \mathbb{Q}(f))$ completely. This completes the proof of (14).

REFERENCES

[1] HADANO, T—KITAOKA, Y.—KUBOTA, T.—NOZAKI M.: *Densities of sets of primes related to decimal expansion of rational numbers* In: Number Theory: Tradition and Modernization (W. Zhang and Y. Tanigawa, eds.), Springer Science + Business Media, Inc. 2006, pp. 67–80.

[2] KITAOKA, Y.: *A statistical relation of roots of a polynomial in different local fields*, Math. of Comp. **78**(2009), 523–536.

[3] KITAOKA Y.: *A statistical relation of roots of a polynomial in different local fields II*, Number Theory : Dreaming in Dreams (Series on Number Theory and Its Application Vol. 6), World Scientific, 2010. pp. 106–126.

[4] KITAOKA, Y.: *A statistical relation of roots of a polynomial in different local fields III*, Osaka J. Math. **49** (2012), 393–420.

[5] KITAOKA, Y.: *A statistical relation of roots of a polynomial in different local fields IV*. Uniform Distribution Theory **8** (2013), no.1, 17–30

Received June 30, 2014
Accepted February 2, 2015

Yoshiyuki Kitaoka
Uzunawa 1085-10, Asahi-cho, Mie, 510-8104
JAPAN
E-mail: kitaoka@meijo-u.ac.jp

Uniform Distribution Theory **10** (2015), no.2, 11–20

uniform distribution theory

TRACTABILITY OF MONTE CARLO INTEGRATION IN HERMITE SPACES

Christian Irrgeher

Dedicated to Professor Harald Niederreiter on the occasion of his 70th birthday

ABSTRACT. We consider multivariate integration in the randomized setting. The function spaces which we study are defined on $\mathbb{R}^s$ with respect to the Gaussian measure and the functions are characterized by the decay of their Hermite coefficients. We study tractability of Monte Carlo integration and give necessary and sufficient conditions to achieve tractability.

Communicated by Henryk Woźniakowski

1. Introduction

Study tractability of multivariate problems, like integration, in reproducing kernel Hilbert spaces goes back to the works of Hickernell [4] and Sloan and Woźniakowski [12]. Since then different notions of tractability were studied for multivariate problems in various function spaces. However, there are only a few results about tractability of multivariate integration of functions defined on unbounded domains, see e.g., [7, 8, 15]. In this paper we want to consider tractability of integration in spaces of functions defined on $\mathbb{R}^s$. For that, we consider the problem of approximating integrals of the form

$$I_s(f) = \int_{\mathbb{R}^s} f(\boldsymbol{x})\varphi_s(\boldsymbol{x})d\boldsymbol{x}, \tag{1}$$

where φ_s denotes the density of the s-dimensional standard Gaussian measure,

$$\varphi_s(\boldsymbol{x}) = \frac{1}{(2\pi)^{s/2}} \exp\left(-\frac{\boldsymbol{x}\cdot\boldsymbol{x}}{2}\right),$$

2010 Mathematics Subject Classification: 65Y20, 65C05.
Keywords: Monte Carlo integration, Tractability, Hermite space.
The author is supported by the Austrian Science Fund (FWF): Project F5509-N26, which is a part of the Special Research Program "Quasi-Monte Carlo Methods: Theory and Applications".

where "$\cdot$" denotes the standard inner product on $\mathbb{R}^s$. Moreover, we consider integrands f which belong to a reproducing kernel Hilbert space $\mathcal{H}(K)$ with norm $\|\cdot\|_K$.

In [5] reproducing kernel Hilbert spaces, so-called Hermite spaces, are introduced for which the problems (1) are well-defined. These function spaces are defined on the $\mathbb{R}^s$ with respect to the Gaussian measure and they are based on Hermite polynomials. For Hermite spaces of functions with polynomially and exponentially decaying Hermite coefficients tractability of multivariate integration was already studied in the worst case setting, see [5] and [6].

In this paper we are interested in approximations of (1) obtained by Monte Carlo (MC) integration rules which are randomized linear algorithms with equal weights and randomly chosen integration nodes. That is, we study tractability in the randomized setting, for more details see Chapter 7 in [10]. We will proceed as in [13] and [3] where tractability of MC integration is studied for the Korobov space and for the Walsh space, respectively.

The rest of the paper is structured as follows: In Section 2 we introduce the general concept of Hermite spaces. Moreover, we present two interesting classes of Hermite spaces: Hermite spaces of functions with polynomially decaying Hermite coefficients and Hermite spaces of functions with exponentially decaying Hermite coefficients. Section 3 deals with tractability of MC integration in Hermite spaces and we give necessary and sufficient conditions to achieve different notions of tractability.

2. The Hermite spaces

We start by introducing some basic facts on *Hermite polynomials*. For more details on Hermite polynomials we refer to [2, 11, 14]. For $k \in \mathbb{N}_0$ the kth Hermite polynomial is given by

$$H_k(x) = \frac{(-1)^k}{\sqrt{k!}} \exp(x^2/2) \frac{d^k}{dx^k} \exp(-x^2/2),$$

where we follow the definition given in [2]. We remark that there are slightly different ways to introduce Hermite polynomials, see, e.g., [14]. For

$$s \geq 2, \quad \boldsymbol{k} = (k_1, \ldots, k_s) \in \mathbb{N}_0^s, \quad \text{and} \quad \boldsymbol{x} = (x_1, \ldots, x_s) \in \mathbb{R}^s$$

we define s-dimensional Hermite polynomials by

$$H_{\boldsymbol{k}}(\boldsymbol{x}) = \prod_{j=1}^{s} H_{k_j}(x_j).$$

It is well-known, see [2], that the Hermite polynomials $\{H_{\boldsymbol{k}}(\boldsymbol{x})\}_{\boldsymbol{k}\in\mathbb{N}_0^s}$ form an orthonormal basis of the space $L^2(\mathbb{R}^s,\varphi_s)$ of function which are square-integrable with respect to the Gaussian measure.

Now we are going to define function spaces based on Hermite polynomials. These kind of function spaces were first introduced in [5]. Let $r:\mathbb{N}_0^s\to\mathbb{R}^+$ be a summable function, i.e., $\sum_{\boldsymbol{k}\in\mathbb{N}_0^s} r(\boldsymbol{k})<\infty$. Define a kernel function

$$K_r(\boldsymbol{x},\boldsymbol{y})=\sum_{\boldsymbol{k}\in\mathbb{N}_0^s} r(\boldsymbol{k})H_{\boldsymbol{k}}(\boldsymbol{x})H_{\boldsymbol{k}}(\boldsymbol{y}) \qquad \text{for} \quad \boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^s$$

and an inner product

$$\langle f,g\rangle_{K_r}=\sum_{\boldsymbol{k}\in\mathbb{N}_0^s}\frac{1}{r(\boldsymbol{k})}\widehat{f}(\boldsymbol{k})\widehat{g}(\boldsymbol{k}),$$

where $\widehat{f}(\boldsymbol{k})=\int_{\mathbb{R}^s} f(\boldsymbol{x})H_{\boldsymbol{k}}(\boldsymbol{x})\varphi_s(\boldsymbol{x})d\boldsymbol{x}$ is the $\boldsymbol{k}$th *Hermite coefficient* of f. Since K_r is symmetric and positive semi-definite, we indeed have that K_r is a reproducing kernel, see, e.g. [3, Chapter 2.3]. Let us denote by $\mathcal{H}(K_r)$ the reproducing kernel Hilbert space corresponding to K_r. The function space $\mathcal{H}(K_r)$ is called a *Hermite space* and the norm in $\mathcal{H}(K_r)$ is defined in the natural way by $\|f\|_{K_r}^2=\langle f,f\rangle_{K_r}$. More details on reproducing kernel Hilbert spaces can be found in [1].

Note that a Hermite space $\mathcal{H}(K_r)$ is fully specified by the function r which regulates the decay of the Hermite coefficients of the functions belonging to $\mathcal{H}(K_r)$. Roughly speaking, the faster r decreases as $\boldsymbol{k}$ grows, the faster the Hermite coefficients of the elements of $\mathcal{H}(K_r)$ decrease.

In this paper we deal with two important classes of Hermite spaces, namely Hermite spaces of functions with polynomially decaying Hermite coefficients and Hermite spaces of functions with exponentially decaying Hermite coefficients. Moreover, we introduce weights to the norm of these function spaces to control the influence of each coordinate.

2.1. Hermite spaces of finite smoothness

To define our function r, we first choose a weight sequence of positive real numbers, $\boldsymbol{\gamma}=\{\gamma_j\}_{j\in\mathbb{N}}$ with $\gamma_j>0$, where we assume that

$$\gamma_1\geq\gamma_2\geq\gamma_3\geq\cdots \tag{2}$$

Furthermore, we fix a parameter $\alpha \in (1, \infty)$. For $k \in \mathbb{N}_0$ we consider

$$r_{\alpha,\gamma_j}(k) = \begin{cases} 1 & \text{if } k = 0, \\ \gamma_j k^{-\alpha} & \text{if } k \neq 0. \end{cases}$$

For a vector $\boldsymbol{k} = (k_1, \ldots, k_s) \in \mathbb{N}_0^s$ we consider

$$r_{s,\alpha,\boldsymbol{\gamma}}(\boldsymbol{k}) = \prod_{j=1}^{s} r_{\alpha,\gamma_j}(k_j).$$

Clearly, it holds that $r_{s,\alpha,\boldsymbol{\gamma}}$ is summable. From now on, we use the following notation for the kernel function,

$$K_{s,\alpha,\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{y}) := \sum_{\boldsymbol{k} \in \mathbb{N}_0^s} r_{s,\alpha,\boldsymbol{\gamma}}(\boldsymbol{k}) H_{\boldsymbol{k}}(\boldsymbol{x}) H_{\boldsymbol{k}}(\boldsymbol{y}),$$

to stress that the reproducing kernel depends on α as well as on the weight sequence $\boldsymbol{\gamma}$. The corresponding reproducing kernel Hilbert space is then given by $\mathcal{H}(K_{s,\alpha,\boldsymbol{\gamma}})$. This choice of r now decreases polynomially fast as $\boldsymbol{k}$ grows, which influences the smoothness of the elements in $\mathcal{H}(K_{s,\alpha,\boldsymbol{\gamma}})$. In [5] it is shown that the smoothness parameter α is related to the differentiability of the functions which makes it reasonable to call $\mathcal{H}(K_{s,\alpha,\boldsymbol{\gamma}})$ a Hermite space of finite smoothness.

2.2. Hermite spaces of analytic functions

Let $\boldsymbol{a} = \{a_j\}_{j \in \mathbb{N}}$ and $\boldsymbol{b} = \{b_j\}_{j \in \mathbb{N}}$ be two weight sequences of real numbers, where we assume that $a_0 := \inf_j a_j > 0$ and $b_0 := \inf_j b_j \geq 1$. Moreover, we fix an $\omega \in (0, 1)$ and for $\boldsymbol{k} \in \mathbb{N}_0^s$ we define

$$r_{s,\omega,\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{k}) = \omega^{|\boldsymbol{k}|_{\boldsymbol{a},\boldsymbol{b}}} := \omega^{\sum_{j=1}^{s} a_j k_j^{\,b_j}} = \prod_{j=1}^{s} \omega^{a_j k_j^{\,b_j}}. \tag{3}$$

We denote the reproducing kernel function by

$$K_{s,\omega,\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{k} \in \mathbb{N}_0^d} \omega^{|\boldsymbol{k}|_{\boldsymbol{a},\boldsymbol{b}}} H_{\boldsymbol{k}}(\boldsymbol{x}) H_{\boldsymbol{k}}(\boldsymbol{y})$$

to indicate again the dependence on the weights. The corresponding Hermite space is then given by $\mathcal{H}(K_{s,\omega,\boldsymbol{a},\boldsymbol{b}})$. With the choice of $r_{s,\omega,\boldsymbol{a},\boldsymbol{b}}$ it follows that the functions in $\mathcal{H}(K_{s,\omega,\boldsymbol{a},\boldsymbol{b}})$ have exponentially decaying Hermite coefficients. Furthermore, this exponential decay guarantees that the functions are extremely smooth, in fact analytic, see [6].

3. Tractability of Monte Carlo integration

Now we study Monte Carlo integration in a Hermite space $\mathcal{H}(K_r)$. For that we consider MC integration rules which are randomized linear algorithms of the form

$$\mathrm{MC}_{n,s}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n; f) = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i),$$

with independent and standard normal distributed random variables $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$. In this setting we are interested in the randomized error of a MC algorithm which is given by

$$e^{\mathrm{MC}}(n, s) = \sup_{f \in \mathcal{H}(K_r), \|f\|_{K_r} \leq 1} \mathbb{E}\left(|I_s(f) - \mathrm{MC}_{n,s}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n; f)|^2 \right)^{\frac{1}{2}},$$

where the expectation is taken with respect to independent and identically distributed random variables $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$. Furthermore, we consider the minimal number of function evaluations which is needed to reduce the initial error by a factor of $\varepsilon \in (0, 1)$, i.e.,

$$n^{\mathrm{MC}}(\varepsilon, s) = \min\{n : e^{\mathrm{MC}}(n, s) \leq \varepsilon\}.$$

Note that the initial error is 1. We want to know how $n^{\mathrm{MC}}(\varepsilon, s)$ depends on ε^{-1} and s. For that we study the tractability of MC algorithms where we follow the notions given in [10]. We say that we have:

(1) *Weak MC-tractability* if

$$\lim_{s+\varepsilon^{-1} \to \infty} \frac{\log(n^{\mathrm{MC}}(\varepsilon, s))}{s + \varepsilon^{-1}} = 0.$$

(2) *Polynomial MC-tractability* if there exist $c, p, q \in \mathbb{R}^+$ such that

$$n^{\mathrm{MC}}(\varepsilon, s) \leq c\, s^q\, \varepsilon^{-p} \qquad \text{for all} \quad s \in \mathbb{N},\ \varepsilon \in (0, 1).$$

(3) *Strong polynomial MC-tractability* if there exist $c, p \in \mathbb{R}^+$ such that

$$n^{\mathrm{MC}}(\varepsilon, s) \leq c\, \varepsilon^{-p} \qquad \text{for all} \quad s \in \mathbb{N},\ \varepsilon \in (0, 1).$$

The infimum of p for which strong polynomial MC-tractability holds is called ε-exponent.

With weak MC-tractability we rule out that the smallest number of function evaluations needed to achieve an ε-approximation depends exponentially on ε^{-1} and s. Polynomial MC-tractability means that $n^{\mathrm{MC}}(\varepsilon, s)$ is bounded polynomially in ε^{-1} and s. In the case of strong polynomial MC-tractability the upper bound is a polynomial in ε^{-1} and independent of the dimension s.

First we derive a formula for the randomized error where we see that the error depends on the number of integration nodes by a factor of $1/\sqrt{n}$. This coincides with the convergence rate of MC algorithms of $\mathcal{O}(n^{1/2})$.

Theorem 1. *For the randomized error of MC integration in the Hermite space $\mathcal{H}(K_r)$ it holds that*

$$e^{\mathrm{MC}}(n,s) = \frac{1}{\sqrt{n}} \left(\max_{\boldsymbol{k} \in \mathbb{N}_0^s \setminus \{\mathbf{0}\}} r(\boldsymbol{k}) \right)^{\frac{1}{2}}.$$

P r o o f. We know for the randomized error that

$$e^{\mathrm{MC}}(n,s) = \frac{1}{\sqrt{n}} \sup_{f \in \mathcal{H}(K_r), \|f\|_{K_r} \leq 1} \left(I_s(f^2) - I_s(f)^2 \right)^{\frac{1}{2}},$$

see, e. g., [9, Theorem 1.1]. Moreover, by Parseval's identity,

$$I_s(f^2) = \int_{\mathbb{R}^s} f(\boldsymbol{x})^2 \varphi_s(\boldsymbol{x}) d\boldsymbol{x} = \sum_{\boldsymbol{k} \in \mathbb{N}_0^s} \hat{f}(\boldsymbol{k})^2$$

and

$$I_s(f) = \int_{\mathbb{R}^d} f(\boldsymbol{x}) \varphi_s(\boldsymbol{x}) d\boldsymbol{x} = \widehat{f}(\mathbf{0}).$$

Hence,

$$\begin{aligned} I_s(f^2) - I_s(f)^2 &= \sum_{\boldsymbol{k} \in \mathbb{N}_0^s \setminus \{\mathbf{0}\}} \hat{f}(\boldsymbol{k})^2 \\ &= \sum_{\boldsymbol{k} \in \mathbb{N}_0^s \setminus \{\mathbf{0}\}} r(\boldsymbol{k})^{-1} \hat{f}(\boldsymbol{k})^2 \, r(\boldsymbol{k}) \\ &\leq \|f\|_{K_r}^2 \max_{\boldsymbol{k} \in \mathbb{N}_0^s \setminus \{\mathbf{0}\}} r(\boldsymbol{k}). \end{aligned} \tag{4}$$

Now we set $\boldsymbol{k}^* = \arg\max_{\boldsymbol{k} \in \mathbb{N}_0^s \setminus \{\mathbf{0}\}} r(\boldsymbol{k})$ and we consider the special integrand $f(\boldsymbol{x}) = H_{\boldsymbol{k}^*}(\boldsymbol{x})$. Then we get for the $\boldsymbol{k}$-th Hermite coefficient of f,

$$\hat{f}(\boldsymbol{k}) = \begin{cases} 1 & \text{if } \boldsymbol{k} = \boldsymbol{k}^*, \\ 0 & \text{otherwise}. \end{cases}$$

Thus, we have that (4) is fulfilled with equality for this choice of f. Hence, it follows that

$$e^{\mathrm{MC}}(n,s) = \frac{1}{\sqrt{n}} \left(\max_{\boldsymbol{k} \in \mathbb{N}_0^s \setminus \{\mathbf{0}\}} r(\boldsymbol{k}) \right)^{\frac{1}{2}}.$$

□

3.1. Tractability in Hermite spaces of finite smoothness

Now we consider MC-tractability of multivariate integration in Hermite spaces $\mathcal{H}(K_{s,\alpha,\boldsymbol{\gamma}})$ of functions of finite smoothness. Since

$$\max_{\boldsymbol{k}\in\mathbb{N}_0^s\setminus\{\mathbf{0}\}} r_{s,\alpha,\boldsymbol{\gamma}}(\boldsymbol{k}) = \max_{k=1,\ldots,s} \prod_{j=1}^{k} \gamma_j,$$

we get from Theorem 1 that the randomized error of MC integration is given by

$$e^{\mathrm{MC}}(n,s) = \frac{1}{\sqrt{n}} \left(\max_{k=1,\ldots,s} \prod_{j=1}^{k} \gamma_j \right)^{\frac{1}{2}}. \tag{5}$$

We remark that this result is similar to the result in [13] and therefore we proceed in the same way to study MC-tractability for the integration problem.

From (5) we see that MC integration is strongly polynomially MC-tractable if and only if $\sup_{s\in\mathbb{N}} \prod_{j=1}^{s} \gamma_j < \infty$. Assume there exists a j with $\gamma_j < 1$, then we have that $\gamma_i < 1$ for all $i \geq j$. Now let j_0 the smallest index such that $\gamma_{j_0} < 1$. Then $\sup_{s\in\mathbb{N}} \prod_{j=1}^{s} \gamma_j < \infty$ is equivalent to $\prod_{j=1}^{j_0-1} \gamma_j < \infty$. On the other hand, if $\gamma_j \geq 1$ for all $j \in \mathbb{N}$, we have that $\sup_{s\in\mathbb{N}} \prod_{j=1}^{s} \gamma_j < \infty$ iff $\prod_{j=1}^{\infty} \gamma_j < \infty$. Altogether, we have that $\sup_{s\in\mathbb{N}} \prod_{j=1}^{s} \gamma_j < \infty$ is equivalent to $\prod_{j=1}^{\infty} \max(\gamma_j, 1) < \infty$ which, in turn, is equivalent to $\sum_{j=1}^{\infty} \max(\log(\gamma_j), 0) < \infty$.

Furthermore, we see from (5) that we have polynomial MC-tractability if and only if there exist $C, q > 0$ such that $\max_{k=1,\ldots,s} \prod_{j=1}^{k} \gamma_j \leq C s^q$. As above, we get that this is equivalent to $\sup_{s\in\mathbb{N}} \sum_{j=1}^{s} \max(\log(\gamma_j), 0))/\log(s) < \infty$.

Finally, we again conclude from (5) that integration is weakly MC-tratable if and only if $\max_{k=1,\ldots,s} \sum_{j=1}^{k} \log(\gamma_j)/s$ approaches zero as s goes to ∞. Again this is equivalent to $\lim_{s\to\infty} \sum_{j=1}^{\infty} \max(\log(\gamma_j), 0)/s = 0$. Now we summarize our results in the next theorem.

Theorem 2. *MC integration in the weighted Hermite space $\mathcal{H}(K_{s,\alpha,\boldsymbol{\gamma}})$ is*

(1) *strongly polynomially MC-tractable iff* $\sum_{j=1}^{\infty} \max(\log(\gamma_j), 0)) < \infty$,

(2) *polynomially MC-tractable iff* $A := \limsup_{s\to\infty} \sum_{j=1}^{s} \frac{\max(\log(\gamma_j),0)}{\log(s)} < \infty$,

(3) *weakly MC-tractable iff* $\lim_{s\to\infty} \sum_{j=1}^{s} \frac{\max(\log(\gamma_j),0)}{s} = 0$.

Let us give some remarks on Theorem 2. We see that the conditions are necessary and sufficient. Moreover, these conditions are fulfilled, if the weight sequence contains weights which are smaller or equal to 1. Especially, in the case

of the unweighted Hermite space, i. e., $\gamma_j = 1$ for all $j \in \mathbb{N}$, we can achieve these three notions of MC-tractability using randomized linear algorithm.

Note that, if we have strong polynomial MC-tractability, then the ε-exponent is 2. Furthermore, the minimal number $n^{\mathrm{MC}}(\varepsilon, s)$ of function evaluations which is needed to guarantee that the randomized error is smaller than ε is

$$\sup_{s\in\mathbb{N}} n^{\mathrm{MC}}(\varepsilon, s) = C\varepsilon^{-2}$$

with $C = \sup_{s\in\mathbb{N}} \prod_{j=1}^{s} \gamma_j < \infty$. If we have polynomial MC-tractability, then

$$n^{\mathrm{MC}}(\varepsilon, s) \leq s^{A+o(1)}\varepsilon^{-2} \qquad \text{as } s \longrightarrow \infty$$

with A as in Theorem 2. Furthermore, we remark that the conditions on MC--tractability of multivariate integration in Hermite spaces of finite smoothness are the same as for Monte Carlo integration in Korobov spaces, see [13], and in Walsh spaces, see [3].

3.2. Tractability in Hermite spaces of analytic functions

For the Hermite space $\mathcal{H}(K_{s,\omega,\boldsymbol{a},\boldsymbol{b}})$ of analytic functions we have that

$$\max_{\boldsymbol{k}\in\mathbb{N}_0^s\setminus\{\boldsymbol{0}\}} r_{s,\omega,\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{k}) = \max_{\boldsymbol{k}\in\mathbb{N}_0^s\setminus\{\boldsymbol{0}\}} \prod_{j=1}^{d} \omega^{a_j {k_j}^{b_j}} = \omega^{a_0} < \infty,$$

because $a_0 = \inf_j a_j > 0$ and $b_j \geq 1$ for all $j \in \mathbb{N}$. From Theorem 1 we get that

$$e^{\mathrm{MC}}(n, s) = \frac{\omega^{a_0}}{\sqrt{n}} \tag{6}$$

and it is easy to see that we can achieve MC-tractability independent of the choice of the weight sequences $\boldsymbol{a}$ and $\boldsymbol{b}$.

Theorem 3. *MC integration in the weighted Hermite space $\mathcal{H}(K_{s,\omega,\boldsymbol{a},\boldsymbol{b}})$ is strongly polynomially MC-tractable, polynomially MC-tractable and weakly MC--tractable for all $\boldsymbol{a}$ and $\boldsymbol{b}$.*

In the worst case setting it is natural to expect exponential convergence for studying multivariate integration in the Hermite space of analytic functions, see [6]. From Theorem 1 it follows that we can not achieve exponential convergence for the error in the randomized setting by using standard Monte Carlo integration, but maybe it could be done by more sophisticated randomized algorithms.

Furthermore, in [6] notions of tractability are considered to study the dependence of n^{MC} on s and $\log \varepsilon^{-1}$. If

$$\lim_{s+\varepsilon^{-1}\to\infty} \frac{\log(n^{\mathrm{MC}}(\varepsilon, s))}{s + \log \varepsilon^{-1}} = 0 \tag{7}$$

with $\log 0 = 0$ taken by convention, it is ruled out that the minimal number n^{MC} of function evaluations to achieve an ε-approximation to the initial error depends exponentially on s and $\log \varepsilon^{-1}$. However, (7) cannot hold for Monte Carlo integration, because

$$n^{\mathrm{MC}}(\varepsilon, s) = \left\lceil \varepsilon^{-2} \omega^{2a_0} \right\rceil .$$

We remark that it is possible in the worst case setting to achieve better convergence rates and related notions of tractability, if we restrict ourself to function spaces of analytic functions, see [6]. However, this is not possible in the randomized setting using Monte Carlo integration as we have seen in this section.

REFERENCES

[1] ARONSZAJN, N.: *Theory of reproducing kernels*, Trans. Amer. Math. Soc. **68** (1950), 337–404.

[2] BOGACHEV, V. I.: *Gaussian Measures*. In: Mathematical Surveys and Monographs. Vol. 62, American Mathematical Society, Providence, 1998.

[3] DICK, J.—PILLICHSHAMMER, F.: *Multivariate integration in weighted Hilbert spaces based on Walsh functions and weighted Sobolev spaces*, J. Complexity **21** (2005), 149–195.

[4] HICKERNELL, F. J.: *Quadrature error bounds with applications to lattice rules*, SIAM J. Numer. Anal. **34** (1997), 853–866.

[5] IRRGEHER, C.—LEOBACHER, G.: *High-dimensional integration on $\mathbb{R}^d$, weighted Hermite spaces, and orthogonal transforms*, J. Complexity **31** (2015), 174–205.

[6] IRRGEHER, C.—KRITZER, P.—LEOBACHER, G.—PILLICHSHAMMER, F.: *Integration in Hermite spaces of analytic functions*, J. Complexity, (2015) (to appear).

[7] KUO, F. Y.—WASILKOWSKI, G. W.—WATERHOUSE, B. J.: Randomly shifted lattice rules for unbounded integrands, J. Complexity **22** (2006), 630–651.

[8] NICHOLS, J. A.—KUO, F. Y.: *Fast CBC construction of randomly shifted lattice rules achieving $\mathcal{O}(n^{-1+\delta})$ convergence for unbounded integrands over $\mathbb{R}^s$ in weighted spaces with POD weights*, J. Complexity **30** (2014), 444–468.

[9] NIEDERREITER, H.: *Random Number Generation and quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania 1992, 1992.

[10] NOVAK, E.—WOŹNIAKOWSKI, H.: *Tractability of Multivariate Problems, Volume I: Linear Information*. EMS, Zurich, 2008.

[11] SANSONE, G.: *Orthogonal Functions*. R. E. Krieger Publishing Company, 1977.

[12] SLOAN, I. H.—WOŹNIAKOWSKI, H.: *When are quasi-Monte Carlo algorithms efficient for high-dimensional integrals?* J. Complexity **14** (1998), 1–33.

[13] SLOAN, I. H.—WOŹNIAKOWSKI, H.: *Tractability of Multivariate Integration for Weighted Korobov Classes*, J. Complexity **17** (2001), 697–721.

[14] SZEGŐ, G.: *Orthogonal Polynomials*, fourth ed., Amer. Math. Soc, Providence, RI, 1975.

[15] WASILKOWSKI, G. W.—WOŹNIAKOWSKI, H.: *Tractability of approximation and integration for weighted tensor product problems over unbounded domains*. In: (K.-T. Fang, F.J. Hickernell, H. Niederreiter, eds.), Monte Carlo and Quasi-Monte Carlo Methods 2000. Springer, Berlin, (2002), pp. 497–522,

Received October 22, 2014
Accepted March 9, 2015

Christian Irrgeher
Department of Financial Mathematics and Applied Number Theory
Johannes Kepler University Linz
Altenberger Strasse 69
4040 Linz
AUSTRIA
E-mail: christian.irrgeher@jku.at

Uniform Distribution Theory **10** (2015), no.2, 21–47

uniform distribution theory

A REDUCED FAST COMPONENT-BY-COMPONENT CONSTRUCTION OF LATTICE POINT SETS WITH SMALL WEIGHTED STAR DISCREPANCY

Ralph Kritzinger—Helene Laimer

Dedicated to Professor Harald Niederreiter on the occasion of his 70th birthday

ABSTRACT. The weighted star discrepancy of point sets appears in the weighted Koksma-Hlawka inequality and thus is a measure for the quality of point sets with respect to their performance in quasi-Monte Carlo algorithms. A specific selection of point sets are lattice point sets whose generating vector can be obtained one component at a time such that the resulting lattice point set has a small weighted star discrepancy.

In this paper we consider a reduced fast component-by-component algorithm which significantly reduces the construction cost for such generating vectors provided that the weights decrease fast enough.

Communicated by Josef Dick

1. Introduction

Given an N-element multiset of points $\{\mathbf{x}_0, \dots, \mathbf{x}_{N-1}\} \subseteq [0,1)^s$, we may approximate integrals over the s-dimensional unit cube by a quasi-Monte Carlo (QMC) rule, i. e.,

$$\int_{[0,1]^s} f(\mathbf{x}) \mathrm{d}\,\mathbf{x} \approx \frac{1}{N} \sum_{n=0}^{N-1} f(\mathbf{x}_n).$$

For detailed information on QMC-integration see [4, 11, 12, 14].

2010 Mathematics Subject Classification: 11K06, 11K38, 65D30, 65D32.
Keywords: lattice point sets, weighted star discrepancy, component-by-component algorithm.
R. Kritzinger is supported by the Austrian Science Fund (FWF): Project F5509-N26, which is a part of the Special Research Program "Quasi-Monte Carlo Methods: Theory and Applications".
H. Laimer is supported by the Austrian Science Fund (FWF): Project F5506-N26, which is a part of the Special Research Program "Quasi-Monte Carlo Methods: Theory and Applications".

In 1998 Sloan and Woźniakowski [23] introduced the concept of weighted function spaces where each group of coordinates is equipped with some weight according to its importance. Denote the set $\{1,\dots,s\}$ by $[s]$ and let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{\mathfrak{u}})_{\mathfrak{u}\subseteq[s]}$ be a weight sequence of non-negative real numbers, which model the importance of the projection of the integrands f in the weighted function space onto the variables x_j for $j \in \mathfrak{u}$. A small weight $\boldsymbol{\gamma}_{\mathfrak{u}}$ means that the projection onto the variables in $\mathfrak{u}$ contributes little to the integration problem. In the present work we consider a special choice of weights, so-called product weights $(\gamma_j)_{j\geq 1}$, where $\boldsymbol{\gamma}_{\mathfrak{u}} = \prod_{j\in\mathfrak{u}} \gamma_j$ and $\boldsymbol{\gamma}_{\emptyset} := 1$, and in particular, the weight γ_j is associated with the variable x_j.

In this paper we assume that $\boldsymbol{\gamma} = (\gamma_j)_{j\geq 1}$ is a non-increasing sequence of positive weights with $\gamma_j \leq 1$ and $(\boldsymbol{\gamma}_{\mathfrak{u}})_{\mathfrak{u}\subseteq[s]}$ are the corresponding product weights. Such weights are useful when considering functions whose dependence on successive variables is decreasing.

A particularly important kind of point sets for QMC-integration are so-called lattice point sets, see [12, 14, 22]. They originated independently from Hlawka [8] and Korobov [10]. A lattice point set $P_N(\mathbf{z}) = \{\mathbf{x}_0,\dots,\mathbf{x}_{N-1}\}$ can be constructed with the aid of a generating vector $\mathbf{z}$. For a positive integer $N \geq 2$ and a vector $\mathbf{z} \in \{1,\dots,N-1\}^s$ the corresponding lattice point set is of the form

$$P_N(\mathbf{z}) = \left\{ \left\{ \frac{k}{N}\mathbf{z} \right\} : k = 0,\dots,N-1 \right\}.$$

Here, for real numbers $x \geq 0$ we write $\{x\} = x - \lfloor x \rfloor$ for the fractional part of x. For vectors $\mathbf{x}$ we apply $\{.\}$ componentwise.

We want to measure the quality of lattice point sets $P_N(\mathbf{z})$ with respect to their performance in a QMC rule. To this end we define the weighted star discrepancy.

Definition 1. Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{\mathfrak{u}})_{\mathfrak{u}\subseteq[s]}$ be a weight sequence and let

$$P_N = \{\mathbf{x}_0,\dots,\mathbf{x}_{N-1}\} \subseteq [0,1]^s \quad \text{be an } N\text{-element point set.}$$

The local discrepancy of the point set P_N at $\mathbf{x} = (x_1,\dots,x_s) \in [0,1]^s$ is defined as

$$\operatorname{discr}(\mathbf{x}, P_N) := \frac{1}{N} \sum_{\mathbf{p}\in P_N} \chi_{[\mathbf{0},\mathbf{x})}(\mathbf{p}) - \prod_{j=1}^{s} x_j,$$

where $\chi_{[\mathbf{0},\mathbf{x})}$ denotes the characteristic function of $[\mathbf{0},\mathbf{x}) := [0,x_1)\times\cdots\times[0,x_s)$. The weighted star discrepancy of P_N is then defined as

$$D^*_{N,\boldsymbol{\gamma}}(P_N) := \sup_{\mathbf{x}\in(0,1]^s} \max_{\emptyset\neq\mathfrak{u}\subseteq[s]} \boldsymbol{\gamma}_{\mathfrak{u}} \left|\operatorname{discr}\big((\mathbf{x}_{\mathfrak{u}},\mathbf{1}), P_N\big)\right|,$$

where $(\mathbf{x}_{\mathfrak{u}},\mathbf{1})$ is the vector $(\tilde{x}_1,\dots,\tilde{x}_s)$ with $\tilde{x}_j = x_j$ if $j \in \mathfrak{u}$ and $\tilde{x}_j = 1$ if $j \notin \mathfrak{u}$.

We denote the weighted star discrepancy of a lattice point set corresponding to some generating vector $\mathbf{z}$ by $D^*_{N,\boldsymbol{\gamma}}(\mathbf{z})$, as $P_N(\mathbf{z})$ is completely determined by $\mathbf{z}$. To see why the weighted star discrepancy is a measure for the quality of our point sets we consider the following identity of Hlawka [7] and Zaremba [24] (see also [4, 12]), given by

$$Q_{N,s}(f) - I_s(f) = \sum_{\emptyset \neq \mathfrak{u} \subseteq [s]} (-1)^{|\mathfrak{u}|} \boldsymbol{\gamma}_{\mathfrak{u}} \int_{[0,1]^{|\mathfrak{u}|}} \operatorname{discr}\big((\mathbf{x}_{\mathfrak{u}}, \mathbf{1}), P_N(\mathbf{z})\big) \boldsymbol{\gamma}_{\mathfrak{u}}^{-1} \frac{\partial^{|\mathfrak{u}|}}{\partial \mathbf{x}_{\mathfrak{u}}} f\,(\mathbf{x}_{\mathfrak{u}}, \mathbf{1})\,\mathrm{d}\mathbf{x}_{\mathfrak{u}},$$

where

$$Q_{N,s}(f) = \frac{1}{N} \sum_{n=0}^{N-1} f(\mathbf{x}_n)$$

denotes the QMC-rule and $I_s(f) = \int_{[0,1]^s} f(\mathbf{x})\,\mathrm{d}\mathbf{x}$ the integral operator.

Applying Hölder's inequality as in [4, 23] for integrals and sums we obtain

$$|Q_{N,s}(f) - I_s(f)| \leq D^*_{N,\boldsymbol{\gamma}}(\mathbf{z}) \|f\|_{\boldsymbol{\gamma}}, \tag{1.1}$$

where $\|.\|_{\boldsymbol{\gamma}}$ is some norm dependent on $\boldsymbol{\gamma}$ but independent of the point set $P_N(\mathbf{z})$.

If f is sufficiently smooth, $\|f\|_{\boldsymbol{\gamma}}$ coincides with the weighted variation of f in the sense of Hardy and Krause. The first factor in (1.1) is the weighted star discrepancy of the point set $P_N(\mathbf{z})$, which depends only on $P_N(\mathbf{z})$ and the weights. Thus we see that the smaller the weighted star discrepancy $D^*_{N,\boldsymbol{\gamma}}(\mathbf{z})$, the better the quality of the lattice point set $P_N(\mathbf{z})$. Hence we want to find lattice point sets $P_N(\mathbf{z})$ with small weighted star discrepancy.

As no explicit constructions for good lattice point sets are known for dimensions $s > 2$, one usually employs computer search algorithms to find good generating vectors. There exist many papers on the construction of generating vectors for lattice point sets with a small weighted star discrepancy: Joe [9] has given a component-by-component construction for generating vectors of lattice point sets with a prime number N of points, which have a weighted star discrepancy of order $N^{-1+\delta}$ for any $\delta > 0$. Their generating vector has a construction cost of order $sN \log N$, where an approach of Nuyens and Cools [19] can be used to reduce the construction cost.

In [20] Joe and Sinescu have achieved the same results for a composite number of lattice points and product weights. Finally in [21] they considered general weights and a prime number of points.

Dick et al. [2] have given a reduced fast algorithm for the construction of generating vectors of lattice point sets with N a prime power. They varied the size of the search space for each coordinate according to its importance and considered the worst-case error of integration in a Korobov space to measure the quality of their lattice point sets.

Let b be an arbitrary prime number and m a positive integer. In the present work we consider lattice point sets with $N = b^m$ elements and study their weighted star discrepancy. As mentioned before, the generating vector $\mathbf{z} = (z_1, \dots, z_s)$ of such lattice point sets can be obtained one component at a time. When using the standard component-by-component construction, in the following frequently abbreviated by CBC construction, each component is chosen from $\{z \in \{1, 2, \dots, b^m - 1\} : \gcd(z, b^m) = 1\}$. As done in [2] for the worst-case error, we speed up the construction of such generating vectors by reducing the search space for each component, while still achieving a small weighted star discrepancy of the corresponding lattice rule. To this end we define non-decreasing $0 \le w_1 \le w_2 \le \cdots \in \mathbb{N}$ and set

$$\mathcal{Z}_{N,w_j} := \begin{cases} \{z \in \{1, 2, \dots, b^{m-w_j} - 1\} : \gcd(z, b^m) = 1\} & \text{if } w_j < m, \\ \{1\} & \text{if } w_j \ge m. \end{cases}$$

Note that these sets have cardinality $b^{m-w_j-1}(b-1)$, for $w_j < m$. In what follows we denote by $\mathcal{Z}^s_{N,\mathbf{w}}$ the Cartesian product $b^{w_1}\mathcal{Z}_{N,w_1} \times \cdots \times b^{w_s}\mathcal{Z}_{N,w_s}$, where $b^{w_j}\mathcal{Z}_{N,w_j}$ means that every element of $\mathcal{Z}_{N,w_j}$ is multiplied by b^{w_j}. We denote by $\mathbf{z} \in \mathcal{Z}^s_{N,\mathbf{w}}$ a vector $\mathbf{z} = (b^{w_1}z_1, \dots, b^{w_s}z_s)$, with $z_j \in \mathcal{Z}_{N,w_j}$ for $j \in [s]$. We study the weighted star discrepancy of lattice point sets $P_N(\mathbf{z})$ with generating vectors $\mathbf{z} \in \mathcal{Z}^s_{N,\mathbf{w}}$. Dick et al. [2] have considered the worst-case error for approximating the integral of functions in suitable spaces by a QMC rule based on lattice point sets. Here, in contrast, we study the weighted star discrepancy of these lattice point sets which is another important quality measure. We will see that for sufficiently fast decreasing weights we can construct lattice point sets with small weighted star discrepancy, while significantly reducing the construction cost in comparison to the standard CBC construction.

It follows from [14, Theorem 3.10 and Theorem 5.6] that

$$D^*_{N,\boldsymbol{\gamma}}(\mathbf{z}) \le \sum_{\mathfrak{u} \subseteq [s]} \gamma_{\mathfrak{u}} \left(1 - \left(1 - \frac{1}{N}\right)^{|\mathfrak{u}|}\right) + \frac{1}{2} R^s_{N,\boldsymbol{\gamma}}(\mathbf{z}), \tag{1.2}$$

where

$$R^s_{N,\boldsymbol{\gamma}}(\mathbf{z}) = \sum_{\mathfrak{u} \subseteq [s]} \gamma_{\mathfrak{u}} R_N(\mathbf{z}, \mathfrak{u}) \tag{1.3}$$

and

$$R_N(\mathbf{z}, \mathfrak{u}) = \frac{1}{N} \sum_{k=0}^{N-1} \prod_{j \in \mathfrak{u}} \left(1 + \sum_{\substack{-\frac{N}{2} < h \le \frac{N}{2} \\ h \ne 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} h k b^{w_j} z_j / N}}{|h|}\right) - 1. \tag{1.4}$$

Using this estimate for the weighted star discrepancy we derive the results in Sections 2, 3 and 4.

Finally, we introduce the concept of tractability [15, 16, 17]. To this end we define the information complexity (often refered to as inverse of the weighted star discrepancy) as

$$N^*(\varepsilon, s) = \min\{N \in \mathbb{N}_0 : D^*_{N,\boldsymbol{\gamma}}(\mathbf{z}) \leq \varepsilon\},$$

which means that $N^*(\varepsilon, s)$ is the minimal number of points required to achieve a weighted star discrepancy of at most ε. Of course we want the information complexity to be as small as possible. Therefore we are interested in how fast it increases when ε^{-1} and s grow. We define the following notions of tractability. We speak of

- polynomial tractability, if there exist constants $C, \tau_1 > 0$ and $\tau_2 \geq 0$ such that

$$N^*(\varepsilon, s) \leq C\varepsilon^{-\tau_1} s^{\tau_2} \quad \text{for all} \quad \varepsilon \in (0,1) \quad \text{and all} \quad s \in \mathbb{N} \quad \text{and of}$$

- strong polynomial tractability, if there exist positive constants C, τ such that

$$N^*(\varepsilon, s) \leq C\varepsilon^{-\tau} \quad \text{for all} \quad \varepsilon \in (0,1) \quad \text{and all} \quad s \in \mathbb{N}.$$

Roughly speaking, a problem is considered tractable if its information complexity's dependence on ε^{-1} and s is not exponential. We will show that the above mentioned reduced fast component-by-component construction finds a generating vector $\mathbf{z}$ of a lattice point set that achieves strong polynomial tractability if

$$\sum_{j=1}^{\infty} \gamma_j b^{w_j} < \infty$$

with a construction cost of

$$O\left(N \log N + \min\{s,t\}N + N \sum_{d=1}^{\min\{s,t\}} (m - w_d) b^{-w_d}\right)$$

operations, where $t = \max\{j \in \mathbb{N} : w_j < m\}$.

The structure of this paper is as follows. In the next section we derive an upper bound for the arithmetic mean of the weighted star discrepancy over all possible lattice point sets constructed by a generating vector $\mathbf{z} \in \mathcal{Z}^s_{N,\mathbf{w}}$. In Sections 3 and 4 we present a reduced fast CBC construction for generating vectors of lattice point sets with small weighted star discrepancy. Finally, in Section 5 we study conditions on the weights γ_j and w_j for achieving strong polynomial tractability.

2. The arithmetic mean over all $\mathbf{z} \in \mathcal{Z}^s_{N,\mathbf{w}}$

First of all we estimate the arithmetic mean of the weighted star discrepancy over all possible generating vectors

$$\mathbf{z} = (b^{w_1} z_1, \dots, b^{w_s} z_s) \in \mathcal{Z}^s_{N,\mathbf{w}},$$

proceeding similarly to [14] and [20]. This yields the existence of a lattice point set with small weighted star discrepancy. The upper bound which we obtain for the arithmetic mean is not the same as for the reduced CBC construction in the next section. Nonetheless, we need large parts of the calculation of the present section to obtain the estimate in Section 3.

Theorem 2.1. *Let $N = b^m$, $(w_j)_{j\geq 1}$ and $\mathcal{Z}^s_{N,\mathbf{w}}$ be as above and let $m \geq 5$. Then there exists a generating vector*

$$\mathbf{z} = (b^{w_1} z_1, \dots, b^{w_s} z_s) \in \mathcal{Z}^s_{N,\mathbf{w}}$$

whose corresponding lattice rule has weighted star discrepancy

$$\begin{aligned} D^*_{N,\boldsymbol{\gamma}}(\mathbf{z}) \leq & \sum_{\mathfrak{u}\subseteq[s]} \gamma_{\mathfrak{u}} \left(1 - \left(1 - \frac{1}{N}\right)^{|\mathfrak{u}|}\right) \\ & + \frac{1}{2}\left(\frac{1}{N}\prod_{j=1}^{s}(\beta_j + \gamma_j S_N)\right. \\ & \left. + \frac{1}{N}\sum_{p=0}^{m-1} b^{m-p-1}(b-1) \prod_{\substack{j=1 \\ w_j \geq m-p}}^{s} (\beta_j + \gamma_j S_N) \prod_{\substack{j=1 \\ w_j < m-p}}^{s} \beta_j - \prod_{j=1}^{s}\beta_j\right), \end{aligned}$$

with $\beta_j = 1 + \gamma_j$ for all $j \in \mathbb{N}$ and

$$S_N = \sum_{\substack{-\frac{N}{2} < h \leq \frac{N}{2} \\ h \neq 0}} \frac{1}{|h|}. \tag{2.1}$$

Remark 1. Provided that the $\gamma_j b^{w_j}$'s are summable, the bound in Theorem 2.1 is of order $N^\delta \log N$ for arbitrary $\delta \in (0,1)$ with an implied constant independent of N and s. Furthermore, note that if all weights $w_j = 0$, then we obtain the result in [20, Theorem 1 and Corollary 1].

P r o o f. As the first sum in (1.2) is independent of $\mathbf{z}$, it is obviously enough to consider the mean

$$M_{N,s,\boldsymbol{\gamma}} := \frac{1}{|\mathcal{Z}^s_{N,\mathbf{w}}|} \sum_{\mathbf{z}\in\mathcal{Z}^s_{N,\mathbf{w}}} R^s_{N,\boldsymbol{\gamma}}(\mathbf{z}) \tag{2.2}$$

of the second sum.

We have from [9, p. 186, eq. 9]

$$\begin{aligned} R^s_{N,\boldsymbol{\gamma}}(\mathbf{z}) &= \frac{1}{N}\sum_{k=0}^{N-1}\prod_{j=1}^{s}\left(\beta_j + \gamma_j \sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_j} z_j/N}}{|h|}\right) - \prod_{j=1}^{s}\beta_j \\ &= \frac{1}{N}\prod_{j=1}^{s}(\beta_j+\gamma_j S_N) + \frac{1}{N}\sum_{k=1}^{N-1}\prod_{j=1}^{s}\left(\beta_j+\gamma_j \sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_j} z_j/N}}{|h|}\right) - \prod_{j=1}^{s}\beta_j. \end{aligned} \tag{2.3}$$

Thus

$$\begin{aligned} M_{N,s,\boldsymbol{\gamma}} &= \frac{1}{N}\prod_{j=1}^{s}(\beta_j+\gamma_j S_N) \\ &\quad + \frac{1}{N}\sum_{k=1}^{N-1}\prod_{j=1}^{s}\left(\frac{1}{|\mathcal{Z}_{N,w_j}|}\sum_{z_j\in\mathcal{Z}_{N,w_j}}\left(\beta_j+\gamma_j \sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_j} z_j/N}}{|h|}\right)\right) - \prod_{j=1}^{s}\beta_j \\ &= \frac{1}{N}\prod_{j=1}^{s}(\beta_j+\gamma_j S_N) \\ &\quad + \frac{1}{N}\sum_{k=1}^{N-1}\prod_{j=1}^{s}\left(\beta_j + \frac{\gamma_j}{|\mathcal{Z}_{N,w_j}|}\sum_{z_j\in\mathcal{Z}_{N,w_j}}\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_j} z_j/N}}{|h|}\right) - \prod_{j=1}^{s}\beta_j. \end{aligned}$$

To avoid lengthy formulas we use the following abbreviations:

$$T_{N,w_j}(k) := \sum_{z_j\in\mathcal{Z}_{N,w_j}}\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_j} z_j/N}}{|h|} \tag{2.4}$$

and

$$L_{N,s,\boldsymbol{\gamma}} := \frac{1}{N}\sum_{k=1}^{N-1}\prod_{j=1}^{s}\left(\beta_j + \frac{\gamma_j}{|\mathcal{Z}_{N,w_j}|}T_{N,w_j}(k)\right). \tag{2.5}$$

Then we have

$$M_{N,s,\boldsymbol{\gamma}} = \frac{1}{N}\prod_{j=1}^{s}(\beta_j + \gamma_j S_N) + L_{N,s,\boldsymbol{\gamma}} - \prod_{j=1}^{s}\beta_j. \tag{2.6}$$

We study $T_{N,w_j}(k)$ distinguishing the two cases $w_j \geq m$ and $w_j < m$.

Case 1 : $w_j \geq m$. This yields $\mathcal{Z}_{N,w_j} = \{1\}$ and thus

$$T_{N,w_j}(k) = \sum_{\substack{-\frac{N}{2}<h\leq\frac{N}{2}\\ h\neq 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_j}/N}}{|h|} = \sum_{\substack{-\frac{N}{2}<h\leq\frac{N}{2}\\ h\neq 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_j-m}}}{|h|} = \sum_{\substack{-\frac{N}{2}<h\leq\frac{N}{2}\\ h\neq 0}} \frac{1}{|h|} = S_N. \tag{2.7}$$

Case 2 : $w_j < m$. Then $\mathcal{Z}_{N,w_j} = \{z \in \{1,2,\dots,b^{m-w_j}-1\} : \gcd(z,N) = 1\}$. According to (2.5) we have to calculate $T_{N,w_j}(k)$ only for $k \in \{1,\dots,b^m-1\}$. We display these k as $k = qb^{m-w_j} + r$ with $q \in \{0,\dots,b^{w_j}-1\}$, $r \in \{0,\dots,b^{m-w_j}-1\}$ and $(q,r) \neq (0,0)$. Then

$$\begin{aligned} T_{N,w_j}(k) &= \sum_{\substack{-\frac{N}{2}<h\leq\frac{N}{2}\\ h\neq 0}} \frac{1}{|h|} \sum_{z_j\in\mathcal{Z}_{N,w_j}} \mathrm{e}^{2\pi \mathrm{i} h(qb^{m-w_j}+r)b^{w_j}z_j/N} \\ &= \sum_{\substack{-\frac{N}{2}<h\leq\frac{N}{2}\\ h\neq 0}} \frac{1}{|h|} \sum_{z_j\in\mathcal{Z}_{N,w_j}} \mathrm{e}^{2\pi \mathrm{i} hqz_j}\mathrm{e}^{2\pi \mathrm{i} hrz_j/b^{m-w_j}} \\ &= \sum_{\substack{-\frac{N}{2}<h\leq\frac{N}{2}\\ h\neq 0}} \frac{1}{|h|} \sum_{z_j\in\mathcal{Z}_{N,w_j}} \mathrm{e}^{2\pi \mathrm{i} hrz_j/b^{m-w_j}}. \end{aligned} \tag{2.8}$$

If $r = 0$, i. e., k a multiple of b^{m-w_j}, this yields

$$T_{N,w_j}(k) = \sum_{\substack{-\frac{N}{2}<h\leq\frac{N}{2}\\ h\neq 0}} \frac{1}{|h|} \sum_{z_j\in\mathcal{Z}_{N,w_j}} 1 = |\mathcal{Z}_{N,w_j}|S_N. \tag{2.9}$$

Next we investigate $r \in \{1,\dots,b^{m-w_j}-1\}$. For any $z_j \in \{1,\dots,b^{m-w_j}-1\}$ we find $\gcd(z_j,N) = \gcd(z_j,b^{m-w_j}) \in \{1,b,b^2,\dots,b^{m-w_j-1}\}$ and hence

$$\sum_{d\mid\gcd(z_j,N)} \mu(d) = \sum_{d\mid\gcd(z_j,b^{m-w_j})} \mu(d) = \begin{cases} 1 & \text{iff } \gcd(z_j,N) = \gcd(z_j,b^{m-w_j}) = 1, \\ 0 & \text{otherwise,} \end{cases}$$

where μ denotes the Möbius function.

For any $z_j \in \{1, \ldots, b^{m-w_j} - 1\}$ this implies $z_j \in \mathcal{Z}_{N,w_j}$ iff $\sum_{d \mid \gcd(z_j, b^{m-w_j})} \mu(d) = 1$.

Inserting this fact into (2.8) we have

$$T_{N,w_j}(k) = \sum_{\substack{-\frac{N}{2} < h \le \frac{N}{2} \\ h \neq 0}} \frac{1}{|h|} \sum_{z_j=1}^{b^{m-w_j}-1} \mathrm{e}^{2\pi \mathrm{i} h r z_j / b^{m-w_j}} \sum_{d \mid \gcd(z_j, b^{m-w_j})} \mu(d). \tag{2.10}$$

Studying the two inner sums we find

$$\begin{aligned} \sum_{z_j=1}^{b^{m-w_j}-1} \mathrm{e}^{2\pi \mathrm{i} h r z_j / b^{m-w_j}} \sum_{d \mid \gcd(z_j, b^{m-w_j})} \mu(d) &= \sum_{d \mid b^{m-w_j}} \mu(d) \sum_{\substack{z_j=1 \\ d \mid z_j}}^{b^{m-w_j}-1} \mathrm{e}^{2\pi \mathrm{i} h r z_j / b^{m-w_j}} \\ &= \sum_{d \mid b^{m-w_j}} \mu(d) \sum_{a=1}^{\frac{b^{m-w_j}}{d}} \mathrm{e}^{2\pi \mathrm{i} h r a d / b^{m-w_j}}, \end{aligned} \tag{2.11}$$

where the latter equality holds since $a \in \left\{1, \ldots, \frac{b^{m-w_j}}{d}\right\}$ yields

$$ad \in \left\{d, 2d, \ldots, b^{m-w_j}\right\} = \left\{1 \le z_j \le b^{m-w_j} - 1 : d | z_j\right\} \cup \left\{b^{m-w_j}\right\}$$

and

$$\sum_{d \mid b^{m-w_j}} \mu(d) = 0,$$

since $w_j < m$.

Changing the order of summation we obtain with (2.11)

$$\begin{aligned} \sum_{z_j=1}^{b^{m-w_j}-1} \mathrm{e}^{2\pi \mathrm{i} h r z_j / b^{m-w_j}} \sum_{d \mid \gcd(z_j, b^{m-w_j})} \mu(d) &= \sum_{d \mid b^{m-w_j}} \mu\left(\frac{b^{m-w_j}}{d}\right) \sum_{a=1}^{d} \mathrm{e}^{2\pi \mathrm{i} h r a / d} \\ &= \sum_{\substack{d \mid b^{m-w_j} \\ d \mid hr}} d\, \mu\left(\frac{b^{m-w_j}}{d}\right). \end{aligned}$$

With (2.10) this leads to

$$T_{N,w_j}(k) = \sum_{\substack{-\frac{N}{2} < h \le \frac{N}{2} \\ h \neq 0}} \frac{1}{|h|} \sum_{\substack{d \mid b^{m-w_j} \\ d \mid hr}} d\, \mu\left(\frac{b^{m-w_j}}{d}\right) = \sum_{d \mid b^{m-w_j}} d\, \mu\left(\frac{b^{m-w_j}}{d}\right) \sum_{\substack{-\frac{N}{2} < h \le \frac{N}{2} \\ h \neq 0 \\ d \mid hr}} \frac{1}{|h|}.$$

Using that $d|hr$ is equivalent to $\frac{d}{\gcd(d,r)}|h$ we display $T_{N,w_j}(k)$ as

$$T_{N,w_j}(k) = \sum_{d|b^{m-w_j}} d\,\mu\left(\frac{b^{m-w_j}}{d}\right) \sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0\\ \frac{d}{\gcd(d,r)}|h}} \frac{1}{|h|}. \tag{2.12}$$

To further investigate $T_{N,w_j}(k)$, we first study sums of the same type as the inner sum in (2.12). For any positive integer a we have

$$\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0\\ a|h}} \frac{1}{|h|} = \sum_{\substack{-\frac{N}{2}<ap\le\frac{N}{2}\\ p\neq 0}} \frac{1}{a|p|} = \frac{1}{a} \sum_{\substack{-\frac{N}{2a}<p\le\frac{N}{2a}\\ p\neq 0}} \frac{1}{|p|} = \frac{1}{a} S_{\frac{N}{a}}, \tag{2.13}$$

where $S_{\frac{N}{a}}$ is defined analogously to (2.1). Combining (2.13) and (2.12) and the property of μ that $\mu(1)=1$, $\mu(b)=-1$ and $\mu(b^i)=0$ for $i\in\mathbb{N}, i\ge 2$ we obtain

$$\begin{aligned}
T_{N,w_j}(k) &= \sum_{d|b^{m-w_j}} d\,\mu\left(\frac{b^{m-w_j}}{d}\right) \frac{\gcd(d,r)}{d} S_{\frac{N}{d}\gcd(d,r)} \\
&= \sum_{d|b^{m-w_j}} \mu\left(\frac{b^{m-w_j}}{d}\right) \gcd(d,r) S_{\frac{N}{d}\gcd(d,r)} \\
&= \sum_{i=0}^{m-w_j} \mu\left(\frac{b^{m-w_j}}{b^i}\right) \gcd(b^i,r) S_{\frac{b^m}{b^i}\gcd(b^i,r)} \\
&= \gcd(b^{m-w_j},r) S_{b^{w_j}\gcd(b^{m-w_j},r)} - \gcd(b^{m-w_j}-1,r) S_{b^{w_j+1}\gcd(b^{m-w_j}-1,r)} \\
&= b^\nu (S_{b^{w_j+\nu}} - S_{b^{w_j+\nu+1}}),
\end{aligned} \tag{2.14}$$

with $\nu\in\{0,\dots,m-w_j-1\}$.

Summarizing, we have for $k\in\{1,\dots,b^m-1\}$

$$T_{N,w_j}(k) = \begin{cases} S_N & \text{if} \quad w_j \ge m, \\ |\mathcal{Z}_{N,w_j}| S_N & \text{if} \quad w_j < m \text{ and } k \equiv 0 \pmod{b^{m-w_j}}, \\ b^\nu (S_{b^{w_j+\nu}} - S_{b^{w_j+\nu+1}}) & \\ \quad \text{with } b^\nu = \gcd(b^{m-w_j},r) & \text{if} \quad w_j < m \text{ and } k \not\equiv 0 \pmod{b^{m-w_j}}. \end{cases} \tag{2.15}$$

Let us choose $t\in\mathbb{N}_0$ such that $w_j < m$ for all $j\le t$ and $w_{t+1}\ge m$. (If $t=0$, then $w_j\ge m$ for all $j\in\mathbb{N}$. In that case we obtain the generating vector $\mathbf{z} = (b^{w_1},\dots,b^{w_s})$.) With this fact we are able to write $L_{N,s,\boldsymbol{\gamma}}$

from formula (2.5) as

$$L_{N,s,\boldsymbol{\gamma}} = \frac{1}{N}\sum_{k=1}^{N-1}\prod_{j=1}^{\min\{t,s\}}\left(\beta_j + \frac{\gamma_j}{|\mathcal{Z}_{N,w_j}|}T_{N,w_j}(k)\right)\prod_{j=t+1}^{s}\left(\beta_j + \frac{\gamma_j}{|\mathcal{Z}_{N,w_j}|}T_{N,w_j}(k)\right)$$
$$= \frac{1}{N}\prod_{j=t+1}^{s}(\beta_j + \gamma_j S_N)\sum_{k=1}^{N-1}\prod_{j=1}^{\min\{t,s\}}\left(\beta_j + \frac{\gamma_j}{|\mathcal{Z}_{N,w_j}|}T_{N,w_j}(k)\right). \tag{2.16}$$

Next we aim at finding bounds for $\frac{T_{N,w_j}(k)}{|\mathcal{Z}_{N,w_j}|}$ for $w_j < m$.

If k is a multiple of b^{m-w_j} we see immediately from (2.15) that

$$\frac{T_{N,w_j}(k)}{|\mathcal{Z}_{N,w_j}|} = \frac{|\mathcal{Z}_{N,w_j}|S_N}{|\mathcal{Z}_{N,w_j}|} = S_N.$$

If k is not a multiple of b^{m-w_j}, we use a formula from Niederreiter [13] for S_n with arbitrary $n \in \mathbb{N}$, given by

$$S_n = 2\log n + 2\gamma - \log 4 + \varepsilon(n), \tag{2.17}$$

where γ denotes the Euler-Mascheroni constant

$$\gamma = \lim_{l\to\infty}\left(\sum_{k=1}^{l}\frac{1}{k} - \log l\right) \approx 0.577216\ldots$$

and

$$\begin{cases} -\frac{4}{n^2} < \varepsilon(n) \le 0, & \text{if } n \text{ is even,} \\ -\frac{3}{n^2} < \varepsilon(n) < \frac{1}{n^2}, & \text{if } n \text{ is odd.} \end{cases} \tag{2.18}$$

From (2.15) we know

$$T_{N,w_j}(k) = b^{\nu}(S_{b^{w_j+\nu}} - S_{b^{w_j+\nu+1}}) < 0. \tag{2.19}$$

With $m \ge 5$ we find $-2 < \frac{T_{N,w_j}(k)}{|\mathcal{Z}_{N,w_j}|} < 0$ for $w_j < m$ and k not a multiple of b^{m-w_j} as follows. The upper bound follows immediately from (2.19). It remains to show the lower bound. First we consider $T_{N,w_j}(k)$ using (2.17). We have

$$\begin{aligned} T_{N,w_j}(k) &= b^{\nu}(S_{b^{w_j+\nu}} - S_{b^{w_j+\nu+1}}) \\ &= b^{\nu}\left(-2\log b + \varepsilon(b^{w_j+\nu}) - \varepsilon(b^{w_j+\nu+1})\right) \\ &= -2b^{\nu}\log b + b^{\nu}\left(\varepsilon(b^{w_j+\nu}) - \varepsilon(b^{w_j+\nu+1})\right). \end{aligned}$$

With (2.18) we obtain

$$\begin{aligned} \left|b^{\nu}\left(\varepsilon(b^{w_j+\nu}) - \varepsilon(b^{w_j+\nu+1})\right)\right| &\le \left|b^{\nu}\left(\varepsilon(b^{w_j+\nu})\right)\right| + \left|b^{\nu}\left(\varepsilon(b^{w_j+\nu+1})\right)\right| \\ &\le 4b^{-2w_j-\nu}\left(1 + \frac{1}{b^2}\right). \end{aligned}$$

Thus

$$\frac{T_{N,w_j}(k)}{|\mathcal{Z}_{N,w_j}|} \geq -\frac{b^{w_j-m+1}}{b-1} 2b^{\nu} \log b - \frac{b^{w_j-m+1}}{b-1} 4b^{-2w_j-\nu} \left(1+\frac{1}{b^2}\right).$$

Recall from (2.15) that $\nu = \log_b \left(\gcd\left(b^{m-w_j}, r\right)\right) \in \{0, 1, \ldots, m-w_j-1\}$. Thus

$$\begin{aligned}\frac{T_{N,w_j}(k)}{|\mathcal{Z}_{N,w_j}|} &\geq -2b^{w_j-m+1+m-w_j-1} \frac{\log b}{b-1} - 4b^{-w_j-m+1-\nu} \frac{1}{b-1}\left(1+\frac{1}{b^2}\right) \\ &\geq -2\frac{\log b}{b-1} - 4b^{-m+1} \frac{1}{b-1}\left(1+\frac{1}{b^2}\right).\end{aligned}$$

Now, with the assumption $m \geq 5$,

$$\begin{aligned}\frac{T_{N,w_j}(k)}{|\mathcal{Z}_{N,w_j}|} &\geq -2\frac{\log b}{b-1} - 4b^{-5+1} \frac{1}{b-1}\left(1+\frac{1}{b^2}\right) \\ &\geq -2\frac{\log 2}{2-1} - 4\cdot 2^{-5+1}\left(1+\frac{1}{2^2}\right) > -2,\end{aligned}$$

and hence

$$-2 < \frac{T_{N,w_j}(k)}{|\mathcal{Z}_{N,w_j}|} < 0 \quad \text{for } w_j < m \quad \text{and} \quad b^{m-w_j} \nmid k.$$

For any integer $p \in \{0, \ldots, m-1\}$ with $b^p \mid k$ and $b^{p+1} \nmid k$ the condition $b^{m-w_j} \nmid k$ is equivalent to $m - w_j > p$ or $w_j < m-p$, respectively. Thus we can display (2.16) as

$$\begin{aligned}L_{N,s,\boldsymbol{\gamma}} &= \frac{1}{N} \prod_{j=t+1}^{s} (\beta_j + \gamma_j S_N) \\ &\quad \times \sum_{p=0}^{m-1} \sum_{\substack{k=1 \\ b^p \mid k \\ b^{p+1} \nmid k}}^{N-1} \prod_{\substack{j=1 \\ w_j \geq m-p}}^{\min\{t,s\}} \left(\beta_j + \frac{\gamma_j}{|\mathcal{Z}_{N,w_j}|} T_{N,w_j}(k)\right) \prod_{\substack{j=1 \\ w_j < m-p}}^{\min\{t,s\}} \left(\beta_j + \frac{\gamma_j}{|\mathcal{Z}_{N,w_j}|} T_{N,w_j}(k)\right) \\ &\leq \frac{1}{N} \prod_{j=t+1}^{s} (\beta_j + \gamma_j S_N) \sum_{p=0}^{m-1} \sum_{\substack{k=1 \\ b^p \mid k \\ b^{p+1} \nmid k}}^{N-1} \prod_{\substack{j=1 \\ w_j \geq m-p}}^{\min\{t,s\}} (\beta_j + \gamma_j S_N) \prod_{\substack{j=1 \\ w_j < m-p}}^{\min\{t,s\}} \beta_j,\end{aligned}$$

where the latter estimate holds since

$$\beta_j > 1, \quad -2 < \frac{T_{N,w_j}(k)}{|\mathcal{Z}_{N,w_j}|} < 0 \quad \text{and} \quad \gamma_j \leq 1.$$

From

$$\begin{aligned}
&\left|\{k \in \{1,\dots,N-1\} : b^p \mid k \quad \text{and} \quad b^{p+1} \nmid k\}\right| \\
&\quad = \left|\{k \in \{1,\dots,b^m-1\} : b^p \mid k\}\right| - \left|\{k \in \{1,\dots,b^m-1\} : b^{p+1} \mid k\}\right| \\
&\quad = b^{m-p} - 1 - \left(b^{m-p-1}-1\right) \\
&\quad = b^{m-p-1}(b-1)
\end{aligned} \tag{2.20}$$

we get

$$L_{N,s,\boldsymbol{\gamma}} \le \frac{1}{N} \prod_{j=t+1}^{s} (\beta_j + \gamma_j S_N) \sum_{p=0}^{m-1} b^{m-p-1}(b-1) \prod_{\substack{j=1 \\ w_j \ge m-p}}^{\min\{t,s\}} (\beta_j + \gamma_j S_N) \prod_{\substack{j=1 \\ w_j < m-p}}^{\min\{t,s\}} \beta_j.$$

Inserting this into (2.6) we obtain for the arithmetic mean

$$\begin{aligned}
M_{N,s,\boldsymbol{\gamma}} = {} & \frac{1}{N} \prod_{j=1}^{s} (\beta_j + \gamma_j S_N) \\
& + \frac{1}{N} \prod_{j=t+1}^{s} (\beta_j + \gamma_j S_N) \sum_{p=0}^{m-1} b^{m-p-1}(b-1) \prod_{\substack{j=1 \\ w_j \ge m-p}}^{\min\{t,s\}} (\beta_j + \gamma_j S_N) \prod_{\substack{j=1 \\ w_j < m-p}}^{\min\{t,s\}} \beta_j \\
& - \prod_{j=1}^{s} \beta_j.
\end{aligned} \tag{2.21}$$

This proves, with (1.2), the existence of a vector $\mathbf{z} \in \mathcal{Z}^s_{N,\mathbf{w}}$ such that the weighted star discrepancy $D^*_{N,\boldsymbol{\gamma}}(\mathbf{z})$ fulfils

$$\begin{aligned}
D^*_{N,\boldsymbol{\gamma}}(\mathbf{z}) &\le \sum_{\mathfrak{u} \subseteq [s]} \gamma_{\mathfrak{u}} \left(1 - \left(1 - \frac{1}{N}\right)^{|\mathfrak{u}|}\right) + \frac{1}{2} \left(\frac{1}{N} \prod_{j=1}^{s} (\beta_j + \gamma_j S_N) \right. \\
&\quad \left. + \frac{1}{N} \prod_{j=t+1}^{s} (\beta_j + \gamma_j S_N) \sum_{p=0}^{m-1} b^{m-p-1}(b-1) \prod_{\substack{j=1 \\ w_j \ge m-p}}^{\min\{t,s\}} (\beta_j + \gamma_j S_N) \prod_{\substack{j=1 \\ w_j < m-p}}^{\min\{t,s\}} \beta_j - \prod_{j=1}^{s} \beta_j \right)
\end{aligned} \tag{2.22}$$

$$\begin{aligned}
&\le \sum_{\mathfrak{u} \subseteq [s]} \gamma_{\mathfrak{u}} \left(1 - \left(1 - \frac{1}{N}\right)^{|\mathfrak{u}|}\right) + \frac{1}{2} \left(\frac{1}{N} \prod_{j=1}^{s} (\beta_j + \gamma_j S_N) \right. \\
&\quad \left. + \frac{1}{N} \sum_{p=0}^{m-1} b^{m-p-1}(b-1) \prod_{\substack{j=1 \\ w_j \ge m-p}}^{s} (\beta_j + \gamma_j S_N) \prod_{\substack{j=1 \\ w_j < m-p}}^{s} \beta_j - \prod_{j=1}^{s} \beta_j \right).
\end{aligned} \tag{2.23}$$

□

3. The reduced CBC construction

In this section we give a component-by-component construction for the generating vector and an upper bound for the weighted star discrepancy of the corresponding lattice rule.

Algorithm 1. Let $N = b^m$ and $(w_j)_{j\geq 1}$ be as above and construct $\mathbf{z} = (b^{w_1} z_1, \ldots, b^{w_s} z_s) \in \mathcal{Z}^s_{N,\mathbf{w}}$ as follows:

(1) Set $z_1 = 1$.

(2) For $d \in [s-1]$ assume $z_1, \ldots, z_d$ to be already found. Choose $z_{d+1} \in \mathcal{Z}_{N,w_{d+1}}$ such that
$$R^{d+1}_{N,\boldsymbol{\gamma}}(b^{w_1} z_1, \ldots, b^{w_d} z_d, b^{w_{d+1}} z)$$
is minimized as a function of z.

(3) Increase d by 1 and repeat the second step until $\mathbf{z} = (b^{w_1} z_1, \ldots, b^{w_s} z_s)$ is found.

In the algorithm above the search space is reduced for each coordinate of $\mathbf{z}$ according to its importance. This results in a considerable reduction of the construction cost as we will see in Section 4. This is why we call this algorithm a reduced CBC-algorithm.

The following theorem gives an upper bound for the figure of merit $R^d_{N,\boldsymbol{\gamma}}$ of lattice point sets with generating vectors obtained from the algorithm above.

Theorem 3.1. *Let $\mathbf{z} = (b^{w_1} z_1, \ldots, b^{w_s} z_s)$ be constructed according to Algorithm 1. Then for every $d \in [s]$,*

$$R^d_{N,\boldsymbol{\gamma}}(b^{w_1} z_1, \ldots, b^{w_d} z_d) \leq \frac{1}{N} \prod_{j=1}^{d} \left(\beta_j + \left(1 + 2b^{\min\{w_j, m\}}\right) \gamma_j S_N\right). \tag{3.1}$$

Corollary 3.2. *Let $N = b^m$ and $(w_j)_{j\geq 1}$ be as above and let*
$$\mathbf{z} = (b^{w_1} z_1, \ldots, b^{w_s} z_s) \in \mathcal{Z}^s_{N,\mathbf{w}}$$
be constructed using Algorithm 1. Then the corresponding lattice point set has a weighted star discrepancy

$$D^*_{N,\boldsymbol{\gamma}}(\mathbf{z}) \leq \sum_{\mathfrak{u}\subseteq[s]} \gamma_{\mathfrak{u}} \left(1 - \left(1 - \frac{1}{N}\right)^{|\mathfrak{u}|}\right) + \frac{1}{2N} \prod_{j=1}^{s} \left(\beta_j + \left(1 + 2b^{\min\{w_j, m\}}\right) \gamma_j S_N\right).$$

Proof. Combining (1.2), (2.1) and Theorem 3.1 we immediately obtain the result. □

To prove Theorem 3.1 we use the the following

Lemma 3.3. *Let $N = b^m$, $(w_j)_{j\geq 1}$ and $\mathcal{Z}_{N,w_j}$ be defined as above and recall from* (2.4) *the notation*

$$T_{N,w_j}(k) \;=\; \sum_{z_j\in\mathcal{Z}_{N,w_j}} \sum_{\substack{-\frac{N}{2}<h\leq\frac{N}{2}\\ h\neq 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} h k b^{w_j} z_j/N}}{|h|}.$$

Then

$$\sum_{k=1}^{N-1} \frac{|T_{N,w_j}(k)|}{|\mathcal{Z}_{N,w_j}|} \leq 2b^{\min\{w_j,m\}} S_N. \tag{3.2}$$

P r o o f. As before, we distinguish the two cases $w_j \geq m$ and $w_j < m$.
Case 1: $w_j \geq m$. Then (2.15) yields

$$\sum_{k=1}^{N-1} \frac{|T_{N,w_j}(k)|}{|\mathcal{Z}_{N,w_j}|} = \sum_{k=1}^{N-1} S_N = (N-1)S_N \leq 2NS_N = 2b^{\min\{w_j,m\}} S_N.$$

Case 2: $w_j < m$. We use (2.15) and (2.8) to find

$$\begin{aligned}\sum_{k=1}^{N-1} \frac{|T_{N,w_j}(k)|}{|\mathcal{Z}_{N,w_j}|} \;&=\; \sum_{\substack{k=1\\ b^{m-w_j}\mid k}}^{N-1} \frac{|T_{N,w_j}(k)|}{|\mathcal{Z}_{N,w_j}|} \;+\; \sum_{\substack{k=1\\ b^{m-w_j}\nmid k}}^{N-1} \frac{|T_{N,w_j}(k)|}{|\mathcal{Z}_{N,w_j}|}\\ &= (b^{w_j}-1)S_N + b^{w_j} \sum_{r=1}^{b^{m-w_j}-1} \frac{|T_{N,w_j}(r)|}{|\mathcal{Z}_{N,w_j}|}.\end{aligned}$$

For any $r \in \{1,\ldots,b^{m-w_j}-1\}$ the condition $\gcd(r, b^{m-w_j}) = b^\nu$ is equivalent to $b^\nu \mid r$ and $b^{\nu+1} \nmid r$ simultaneously. Using this we investigate the last sum in the above equation:

$$\sum_{r=1}^{b^{m-w_j}-1} \frac{|T_{N,w_j}(r)|}{|\mathcal{Z}_{N,w_j}|} = \frac{1}{|\mathcal{Z}_{N,w_j}|} \sum_{\nu=0}^{m-w_j-1} \sum_{\substack{r=1\\ b^\nu\mid r\\ b^{\nu+1}\nmid r}}^{b^{m-w_j}-1} |T_{N,w_j}(r)|.$$

Once more with the aid of (2.15) this yields

$$\begin{aligned}\sum_{r=1}^{b^{m-w_j}-1} \frac{|T_{N,w_j}(r)|}{|\mathcal{Z}_{N,w_j}|} &= \frac{1}{|\mathcal{Z}_{N,w_j}|} \sum_{\nu=0}^{m-w_j-1} \sum_{\substack{r=1\\ b^\nu\mid r\\ b^{\nu+1}\nmid r}}^{b^{m-w_j}-1} \left|b^\nu (S_{b^{w_j+\nu}} - S_{b^{w_j+\nu+1}})\right|\\ &= \frac{1}{|\mathcal{Z}_{N,w_j}|} \sum_{\nu=0}^{m-w_j-1} \sum_{\substack{r=1\\ b^\nu\mid r\\ b^{\nu+1}\nmid r}}^{b^{m-w_j}-1} b^\nu (S_{b^{w_j+\nu+1}} - S_{b^{w_j+\nu}}).\end{aligned}$$

Analogously to (2.20) we find

$$\left|\left\{r \in \left\{1,\dots,b^{m-w_j}-1\right\} : b^\nu \mid r \text{ and } b^{\nu+1} \nmid r\right\}\right| = b^{m-w_j-\nu-1}(b-1)$$

and hence

$$\sum_{r=1}^{b^{m-w_j}-1} \frac{|T_{N,w_j}(r)|}{|\mathcal{Z}_{N,w_j}|} \;=\; \sum_{\nu=0}^{m-w_j-1} \left(S_{b^{w_j+\nu+1}} - S_{b^{w_j+\nu}}\right) \;=\; S_N - S_{b^{w_j}}\,.$$

Altogether we have

$$\begin{aligned}\sum_{k=1}^{N-1} \frac{|T_{N,w_j}(k)|}{|\mathcal{Z}_{N,w_j}|} \;&=\; (b^{w_j}-1)S_N + b^{w_j}(S_N - S_{b^{w_j}}) \\ &\le\; 2b^{w_j}S_N \;=\; 2b^{\min\{w_j,m\}}S_N\end{aligned}$$

and the proof is complete. □

With the aid of Lemma 3.3 we are able to prove Theorem 3.1 using induction on d.

P r o o f. According to Algorithm 1 we set $z_1 = 1$ in Step 1. We have to show that

$$R^1_{N,\boldsymbol{\gamma}}(b^{w_1}) \le \frac{1}{N}\left(\beta_1 + \left(1 + 2b^{\min\{w_1,m\}}\right)\gamma_1 S_N\right).$$

With (2.3) we have

$$\begin{aligned}R^1_{N,\boldsymbol{\gamma}}(b^{w_1}) &= \frac{1}{N}\sum_{k=0}^{N-1}\left(\beta_1 + \gamma_1 \sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\ne 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_1}/N}}{|h|}\right) - \beta_1 \\ &= \frac{1}{N}\sum_{k=0}^{N-1}\gamma_1 \sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\ne 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_1}/N}}{|h|}.\end{aligned}$$

Again, we consider the two cases $w_1 \ge m$ and $w_1 < m$ separately.

Case 1: $w_1 \ge m$. Then

$$\begin{aligned}R^1_{N,\boldsymbol{\gamma}}(b^{w_1}) &= \frac{1}{N}\sum_{k=0}^{N-1}\gamma_1 \sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\ne 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_1-m}}}{|h|} = \frac{1}{N}\gamma_1 N S_N \le \frac{1}{N}(1+\gamma_1+2N\gamma_1 S_N) \\ &= \frac{1}{N}\left(\beta_1 + 2b^{\min\{w_1,m\}}\gamma_1 S_N\right) \le \frac{1}{N}\left(\beta_1 + \left(1+2b^{\min\{w_1,m\}}\right)\gamma_1 S_N\right),\end{aligned}$$

which is the desired result.

Case 2: $w_1 < m$. After interchanging the two sums, once more, we split up the inner sum as follows:

$$R^1_{N,\boldsymbol{\gamma}}(b^{w_1}) = \frac{\gamma_1}{N} \sum_{\substack{-\frac{N}{2} < h \le \frac{N}{2} \\ h \neq 0}} \frac{1}{|h|} \sum_{k=0}^{N-1} \mathrm{e}^{2\pi \mathrm{i} h k / b^{m-w_1}}$$

$$= \frac{\gamma_1}{N} \sum_{\substack{-\frac{N}{2} < h \le \frac{N}{2} \\ h \neq 0 \\ b^{m-w_1} \mid h}} \frac{1}{|h|} \sum_{k=0}^{N-1} \mathrm{e}^{2\pi \mathrm{i} h k / b^{m-w_1}} + \frac{\gamma_1}{N} \sum_{\substack{-\frac{N}{2} < h \le \frac{N}{2} \\ h \neq 0 \\ b^{m-w_1} \nmid h}} \frac{1}{|h|} \sum_{k=0}^{N-1} \mathrm{e}^{2\pi \mathrm{i} h k / b^{m-w_1}}.$$

Now we are able to compute the inner sums. The first one sums to N, whereas the second one equals zero which can immediately be seen by applying the formula for finite geometric series. Thus

$$R^1_{N,\boldsymbol{\gamma}}(b^{w_1}) = \gamma_1 \sum_{\substack{-\frac{N}{2} < h \le \frac{N}{2} \\ h \neq 0 \\ b^{m-w_1} \mid h}} \frac{1}{|h|}.$$

We use (2.13) to find

$$R^1_{N,\boldsymbol{\gamma}}(b^{w_1}) = \gamma_1 \frac{1}{b^{m-w_1}} S_{\frac{N}{b^{m-w_1}}} = \frac{\gamma_1}{N} b^{w_1} S_{b^{w_1}}$$

$$\le \frac{\gamma_1}{N} b^{w_1} S_N \le \frac{1}{N} \left(\beta_1 + 2b^{w_1} \gamma_1 S_N\right)$$

$$\le \frac{1}{N} \left(\beta_1 + \left(1 + 2b^{\min\{w_1, m\}}\right) \gamma_1 S_N \right),$$

as it is claimed.

Let $d \in [s-1]$ and assume that we have some $\mathbf{z} \in \mathcal{Z}^d_{N,\mathbf{w}}$, such that

$$R^d_{N,\boldsymbol{\gamma}}(b^{w_1} z_1, \ldots, b^{w_d} z_d) \le \frac{1}{N} \prod_{j=1}^{d} \left(\beta_j + \left(1 + 2b^{\min\{w_j, m\}}\right) \gamma_j S_N \right).$$

We have to prove the existence of a $z_{d+1} \in \mathcal{Z}_{N,w_{d+1}}$ with

$$R^{d+1}_{N,\boldsymbol{\gamma}}(b^{w_1} z_1, \ldots, b^{w_d} z_d, b^{w_{d+1}} z_{d+1}) \le \frac{1}{N} \prod_{j=1}^{d+1} \left(\beta_j + \left(1 + 2b^{\min\{w_j, m\}}\right) \gamma_j S_N \right).$$

Using again (2.3) we have for any $z_{d+1} \in \mathcal{Z}_{N,w_{d+1}}$ that

$$
\begin{aligned}
&R^{d+1}_{N,\boldsymbol{\gamma}}(b^{w_1}z_1,\dots,b^{w_d}z_d,b^{w_{d+1}}z_{d+1}) \\
&\quad= \frac{1}{N}\sum_{k=0}^{N-1}\prod_{j=1}^{d}\left(\beta_j+\gamma_j\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}}\frac{\mathrm{e}^{2\pi\mathrm{i}hkb^{w_j}z_j/N}}{|h|}\right) \\
&\qquad\times\left(\beta_{d+1}+\gamma_{d+1}\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}}\frac{\mathrm{e}^{2\pi\mathrm{i}hkb^{w_{d+1}}z_{d+1}/N}}{|h|}\right)-\beta_{d+1}\prod_{j=1}^{d}\beta_j \\
&\quad= \beta_{d+1}R^{d}_{N,\boldsymbol{\gamma}}(b^{w_1}z_1,\dots,b^{w_d}z_d) \\
&\qquad+\frac{\gamma_{d+1}}{N}\sum_{k=0}^{N-1}\prod_{j=1}^{d}\left(\beta_j+\gamma_j\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}}\frac{\mathrm{e}^{2\pi\mathrm{i}hkb^{w_j}z_j/N}}{|h|}\right)\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}}\frac{\mathrm{e}^{2\pi\mathrm{i}hkb^{w_{d+1}}z_{d+1}/N}}{|h|} \\
&\quad= \beta_{d+1}R^{d}_{N,\boldsymbol{\gamma}}(b^{w_1}z_1,\dots,b^{w_d}z_d)+\frac{\gamma_{d+1}S_N}{N}\prod_{j=1}^{d}(\beta_j+\gamma_jS_N) \\
&\qquad+\frac{\gamma_{d+1}}{N}\sum_{k=1}^{N-1}\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}}\frac{\mathrm{e}^{2\pi\mathrm{i}hkb^{w_{d+1}}z_{d+1}/N}}{|h|}\prod_{j=1}^{d}\left(\beta_j+\gamma_j\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}}\frac{\mathrm{e}^{2\pi\mathrm{i}hkb^{w_j}z_j/N}}{|h|}\right).
\end{aligned}
\tag{3.3}
$$

Next we consider the arithmetic mean of

$$R^{d+1}_{N,\boldsymbol{\gamma}}(b^{w_1}z_1,\dots,b^{w_d}z_d,b^{w_{d+1}}z) \quad \text{over all} \quad z\in\mathcal{Z}_{N,w_{d+1}}.$$

As only the third summand in (3.3) depends on the $(d+1)$-st coordinate it suffices to investigate the mean of this summand. Clearly, if we have some upper bound for the mean over all

$$z\in\mathcal{Z}_{N,w_{d+1}}, \quad \text{there exists} \quad z_{d+1}\in\mathcal{Z}_{N,w_{d+1}}$$

which satisfies this bound.

Thus we study

$$\frac{1}{|\mathcal{Z}_{N,w_{d+1}}|}\sum_{z\in\mathcal{Z}_{N,w_{d+1}}}\frac{\gamma_{d+1}}{N}\sum_{k=1}^{N-1}\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}}\frac{\mathrm{e}^{2\pi\mathrm{i}hkb^{w_{d+1}}z/N}}{|h|}$$
$$\times\prod_{j=1}^{d}\left(\beta_j+\gamma_j\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}}\frac{\mathrm{e}^{2\pi\mathrm{i}hkb^{w_j}z_j/N}}{|h|}\right).$$

We bound this term by its absolute value:

$$\begin{aligned}
&\left|\frac{1}{|\mathcal{Z}_{N,w_{d+1}}|}\sum_{z\in\mathcal{Z}_{N,w_{d+1}}}\frac{\gamma_{d+1}}{N}\sum_{k=1}^{N-1}\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}}\frac{\mathrm{e}^{2\pi\mathrm{i}hkb^{w_{d+1}}z/N}}{|h|}\right.\\
&\quad\left.\times\prod_{j=1}^{d}\left(\beta_j+\gamma_j\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}}\frac{\mathrm{e}^{2\pi\mathrm{i}hkb^{w_j}z_j/N}}{|h|}\right)\right|\\
&\le\frac{\gamma_{d+1}}{N}\sum_{k=1}^{N-1}\frac{1}{|\mathcal{Z}_{N,w_{d+1}}|}\left|\sum_{z\in\mathcal{Z}_{N,w_{d+1}}}\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}}\frac{\mathrm{e}^{2\pi\mathrm{i}hkb^{w_{d+1}}z/N}}{|h|}\right|\\
&\quad\times\prod_{j=1}^{d}\left(\beta_j+\gamma_j\sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}}\frac{\left|\mathrm{e}^{2\pi\mathrm{i}hkb^{w_j}z_j/N}\right|}{|h|}\right)\\
&\le\frac{\gamma_{d+1}}{N}\sum_{k=1}^{N-1}\frac{|T_{N,w_{d+1}}(k)|}{|\mathcal{Z}_{N,w_{d+1}}|}\prod_{j=1}^{d}(\beta_j+\gamma_j S_N)\\
&\le\frac{\gamma_{d+1}}{N}2b^{\min\{w_{d+1},m\}}S_N\prod_{j=1}^{d}(\beta_j+\gamma_j S_N),
\end{aligned}$$

where the last estimate stems from an application of Lemma 3.3. Combining this with (3.3) we have shown the existence of a $z_{d+1}\in\mathcal{Z}_{N,w_{d+1}}$ such that

$$
\begin{aligned}
&R^{d+1}_{N,\boldsymbol{\gamma}}(b^{w_1}z_1,\dots,b^{w_d}z_d,b^{w_{d+1}}z_{d+1})\\
&\le \beta_{d+1}R^{d}_{N,\boldsymbol{\gamma}}(b^{w_1}z_1,\dots,b^{w_d}z_d)\\
&\quad+\frac{\gamma_{d+1}S_N}{N}\prod_{j=1}^{d}(\beta_j+\gamma_j S_N)\\
&\quad+\frac{\gamma_{d+1}}{N}2b^{\min\{w_{d+1},m\}}S_N\prod_{j=1}^{d}(\beta_j+\gamma_j S_N).
\end{aligned}
$$

We use the induction hypothesis to find

$$
\begin{aligned}
&R^{d+1}_{N,\boldsymbol{\gamma}}(b^{w_1}z_1,\dots,b^{w_d}z_d,b^{w_{d+1}}z_{d+1})\\
&\le \frac{\beta_{d+1}}{N}\prod_{j=1}^{d}\left(\beta_j+\left(1+2b^{\min\{w_j,m\}}\right)\gamma_j S_N\right)\\
&\quad+\frac{\gamma_{d+1}S_N}{N}\prod_{j=1}^{d}(\beta_j+\gamma_j S_N)\left(1+2b^{\min\{w_{d+1},m\}}\right)\\
&\le \left(\beta_{d+1}+\left(1+2b^{\min\{w_{d+1},m\}}\right)\gamma_{d+1}S_N\right)\\
&\quad\times\frac{1}{N}\prod_{j=1}^{d}\left(\beta_j+\left(1+2b^{\min\{w_j,m\}}\right)\gamma_j S_N\right)\\
&=\frac{1}{N}\prod_{j=1}^{d+1}\left(\beta_j+\left(1+2b^{\min\{w_j,m\}}\right)\gamma_j S_N\right),
\end{aligned}
$$

which completes the proof. □

4. The reduced fast CBC construction

By now we have seen how we can construct a generating vector of a lattice point set with low weighted star discrepancy with a reduced CBC construction as in the previous section. Now we study the construction cost of this algorithm. In fact the CBC algorithm given in Section 3 can be made faster to construct generating vectors for relatively large N and s. To show this we follow closely [2] and [12].

Let $d \in [s-1]$ and assume that we have already found $(b^{w_1}z_1, \dots, b^{w_d}z_d)$. Then we have (cf. (2.3))

$$R^d_{N,\boldsymbol{\gamma}}(b^{w_1}z_1, \dots, b^{w_d}z_d) = \frac{1}{N}\sum_{k=0}^{N-1}\prod_{j=1}^{d}\left(\beta_j + \gamma_j \sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_j}z_j/N}}{|h|}\right) - \prod_{j=1}^{d}\beta_j.$$

Define $r(h) = \max\{1, |h|\}$. Then

$$\begin{aligned}\beta_j + \gamma_j \sum_{\substack{-\frac{N}{2}<h\le\frac{N}{2}\\ h\neq 0}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_j}z_j/N}}{|h|} &= \beta_j + \gamma_j \left(\sum_{-\frac{N}{2}<h\le\frac{N}{2}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_j}z_j/N}}{r(h)} - 1\right)\\ &= 1 + \gamma_j \sum_{-\frac{N}{2}<h\le\frac{N}{2}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_j}z_j/N}}{r(h)}.\end{aligned}$$

Hence we have

$$\begin{aligned}R^d_{N,\boldsymbol{\gamma}}(b^{w_1}z_1, \dots, b^{w_d}z_d) &= \frac{1}{N}\sum_{k=0}^{N-1}\prod_{j=1}^{d}\left(1 + \gamma_j \sum_{-\frac{N}{2}<h\le\frac{N}{2}} \frac{\mathrm{e}^{2\pi \mathrm{i} hkb^{w_j}z_j/N}}{r(h)}\right) - \prod_{j=1}^{d}\beta_j\\ &= \frac{1}{N}\sum_{k=0}^{N-1}\eta_d(k) - \prod_{j=1}^{d}\beta_j, \qquad (4.1)\end{aligned}$$

where we have defined

$$\eta_d(k) = \prod_{j=1}^{d}\left(1 + \gamma_j \phi\left(\frac{kb^{w_j}z_j}{N}\right)\right)$$

and

$$\phi(x) = \sum_{-\frac{N}{2}<h\le\frac{N}{2}} \frac{\mathrm{e}^{2\pi \mathrm{i} hx}}{r(h)}.$$

However, this is exactly the situation as dealt with in [12, Section 4.2]. Thus we know that $\phi\left(\frac{kb^{w_j}z_j}{N}\right)$ takes on at most N different values, namely

$$\phi(0), \phi\left(\frac{1}{N}\right), \dots, \phi\left(\frac{N-1}{N}\right),$$

which can be computed in $O(N \log N)$ operations and stored in a memory space of size $O(N)$, as demonstrated in [12, Section 4.2].

Next we investigate one actual step of the CBC construction. Assuming that we have already found $(b^{w_1}z_1, \dots, b^{w_d}z_d) \in \mathcal{Z}^d_{N,\mathbf{w}}$ we have to minimize

$$R^{d+1}_{N,\boldsymbol{\gamma}}(b^{w_1}z_1, \dots, b^{w_d}z_d, b^{w_{d+1}}z)$$

as a function of $z \in \mathcal{Z}_{N,w_{d+1}}$ to find $z_{d+1} \in \mathcal{Z}_{N,w_{d+1}}$. For $w_{d+1} \geq m$ we just set $z_{d+1} = 1$ and we are done. Therefore let $w_{d+1} < m$. Considering (4.1) we have

$$\begin{aligned}
&R^{d+1}_{N,\boldsymbol{\gamma}}(b^{w_1}z_1, \dots, b^{w_d}z_d, b^{w_{d+1}}z_{d+1}) \\
&= \frac{1}{N}\sum_{k=0}^{N-1} \eta_{d+1}(k) - \prod_{j=1}^{d+1}\beta_j \\
&= \frac{1}{N}\sum_{k=0}^{N-1} \eta_d(k)\left(1 + \gamma_{d+1}\phi\left(\frac{kb^{w_{d+1}}z_{d+1}}{N}\right)\right) - \prod_{j=1}^{d+1}\beta_j \\
&= \frac{1}{N}\sum_{k=0}^{N-1} \eta_d(k)\left(1 + \gamma_{d+1}\phi\left(\left\{\frac{kb^{w_{d+1}}z_{d+1}}{N}\right\}\right)\right) - \prod_{j=1}^{d+1}\beta_j.
\end{aligned}$$

It is obviously enough to minimize $\sum_{k=0}^{N-1} \eta_d(k)\phi\left(\left\{\frac{kb^{w_{d+1}}z_{d+1}}{N}\right\}\right)$. To do this we proceed analogously to [2]. We define the matrix

$$A = \left(\phi\left(\left\{\frac{kb^{w_{d+1}}z}{N}\right\}\right)\right)_{z\in\mathcal{Z}_{N,w_{d+1}},k\in\{0,\dots,N-1\}},$$

the vector

$$\boldsymbol{\eta}_d = (\eta_d(0), \eta_d(1), \dots, \eta_d(N-1))^\top$$

and

$$T_d(z) = \sum_{k=0}^{N-1} \eta_d(k)\phi\left(\left\{\frac{kb^{w_{d+1}}z}{N}\right\}\right).$$

Then

$$A\boldsymbol{\eta}_d = \boldsymbol{T}_d(z) := (T_d(z))_{z\in\mathcal{Z}_{N,w_{d+1}}}.$$

We can display the matrix A as

$$A = \left(\Omega^{(m-w_{d+1})}, \dots, \Omega^{(m-w_{d+1})}\right),$$

with

$$\Omega^{(l)} = \left(\phi\left(\left\{\frac{kz}{b^l}\right\}\right)\right)_{z\in\mathcal{Z}_{b^l,0},k\in\{0,\dots,b^l-1\}}.$$

Again analogously to [2] we obtain the following reduced fast CBC algorithm.

Algorithm 2.

a) Compute $\phi\left(\frac{r}{N}\right)$ for all $r = 0, \ldots, N-1$.

b) Set $\eta_1(k) = 1 + \gamma_1 \phi\left(\left\{\frac{k b^{w_1} z_1}{N}\right\}\right)$ for $k = 0, \ldots, N-1$.

c) Set $z_1 = 1$. Set $d = 2$ and recall that we have defined $t = \max\{j : w_j < m\}$. While $d \le \min\{s, t\}$,

1. partition $\boldsymbol{\eta}_{d-1}$ into b^{w_d} vectors $\boldsymbol{\eta}_{d-1}^{(1)}, \ldots, \boldsymbol{\eta}_{d-1}^{(b^{w_d})}$ of length b^{m-w_d} and let $\boldsymbol{\eta}' = \boldsymbol{\eta}_{d-1}^{(1)} + \cdots + \boldsymbol{\eta}_{d-1}^{(b^{w_d})}$ denote their sum,
2. let $T_d(z) = \Omega^{(m-w_d)} \boldsymbol{\eta}'$,
3. let $z_d = \arg\min_z T_d(z)$,
4. let $\eta_d(k) = \eta_{d-1}(k)\left(1 + \gamma_d \phi\left(\left\{\frac{k b^{w_d} z_d}{N}\right\}\right)\right)$ for $k = 0, \ldots, N-1$,
5. increase d by 1.

If $s > t$, then set $z_{t+1} = \cdots = z_s = 1$. Then we have

$$R^s_{N,\boldsymbol{\gamma}}\left(b^{w_1} z_1, \ldots, b^{w_s} z_s\right) = \frac{1}{N} \sum_{k=0}^{N-1} \eta_s(k) - \prod_{j=1}^{s} \beta_j.$$

Using [2, 12, 18, 19] we find that Algorithm 2 has a construction cost of

$$O\left(N \log N + \min\{s,t\} N + N \sum_{d=1}^{\min\{s,t\}} (m - w_d) b^{-w_d}\right)$$

operations, in comparison to $O(sN \log N)$ operations for the standard CBC algorithm used for example in [20].

5. Conditions for strong polynomial tractability

Let $\mathbf{z} = (b^{w_1} z_1, \ldots, b^{w_s} z_s) \in \mathcal{Z}^s_{N,\mathbf{w}}$ be constructed with Algorithm 1 or 2 and consider the corresponding lattice rule. We are interested in conditions for tractability of the weighted star discrepancy of such lattice point sets. From (1.2) and (1.3) we know

$$D^*_{N,\boldsymbol{\gamma}}(\mathbf{z}) \le \sum_{\mathfrak{u} \subseteq [s]} \gamma_{\mathfrak{u}} \left(1 - \left(1 - \frac{1}{N}\right)^{|\mathfrak{u}|}\right) + \frac{1}{2} R^s_{N,\boldsymbol{\gamma}}(\mathbf{z}).$$

For now, let us assume that the $\gamma_j b^{w_j}$'s are summable, i. e.,

$$\sum_{j=1}^{\infty} \gamma_j b^{w_j} < \infty.$$

Similar to Joe and Sinescu in [9] and [20], we see that in this case

$$D^*_{N,\boldsymbol{\gamma}}(\mathbf{z}) \leq \frac{\max\{1,\Gamma\} \exp\left(\sum_{j=1}^{\infty} \gamma_j\right)}{N} + \frac{1}{2} R^s_{N,\boldsymbol{\gamma}}(\mathbf{z}),$$

where

$$\Gamma = \sum_{j=1}^{\infty} \frac{\gamma_j}{1+\gamma_j} < \infty.$$

In particular, considering our assumption that the $\gamma_j b^{w_j}$'s are summable, the constant

$$\max\{1,\Gamma\} \exp\left(\sum_{j=1}^{\infty} \gamma_j\right)$$

is indeed finite.

Theorem 3.1 yields

$$R^s_{N,\boldsymbol{\gamma}}(\mathbf{z}) \leq \frac{1}{N} \prod_{j=1}^{s} \left(\beta_j + \left(1 + 2b^{\min\{w_j,m\}}\right) \gamma_j S_N\right)$$

and hence we have

$$\begin{aligned} D^*_{N,\boldsymbol{\gamma}}(\mathbf{z}) &\leq \frac{1 + \max\{1,\Gamma\} \exp\left(\sum_{j=1}^{\infty} \gamma_j\right)}{N} \prod_{j=1}^{s} \left(\beta_j + \left(1 + 2b^{\min\{w_j,m\}}\right) \gamma_j S_N\right) \\ &= \frac{c_{\boldsymbol{\gamma}}}{N} \prod_{j=1}^{s} \left(\beta_j + \left(1 + 2b^{\min\{w_j,m\}}\right) \gamma_j S_N\right), \end{aligned} \tag{5.1}$$

with $c_{\boldsymbol{\gamma}} = 1 + \max\{1,\Gamma\} \exp\left(\sum_{j=1}^{\infty} \gamma_j\right)$ independent of s. We study the right-hand side of (5.1)

$$\begin{aligned} &\frac{c_{\boldsymbol{\gamma}}}{N} \prod_{j=1}^{s} \left(\beta_j + \left(1 + 2b^{\min\{w_j,m\}}\right) \gamma_j S_N\right) \\ &\leq \frac{c_{\boldsymbol{\gamma}}}{N} \prod_{j=1}^{s} \left(\beta_j + \left(1 + 2b^{\min\{w_j,m\}}\right) \gamma_j 2 \left(\log\left\lfloor \frac{N}{2} \right\rfloor + 1\right)\right) \\ &\leq \frac{c_{\boldsymbol{\gamma}}}{N} \prod_{j=1}^{s} \left(\beta_j + \left(1 + 2b^{\min\{w_j,m\}}\right) \gamma_j 4 \log N\right) \\ &= \frac{c_{\boldsymbol{\gamma}}}{N} \prod_{j=1}^{s} \left(1 + \gamma_j \left(1 + 4\left(1 + 2b^{\min\{w_j,m\}}\right) \log N\right)\right), \end{aligned} \tag{5.2}$$

where we have used

$$S_N = \sum_{\substack{-\frac{N}{2} < h \le \frac{N}{2} \\ h \ne 0}} \frac{1}{|h|} \le 2 \sum_{h=1}^{\lfloor \frac{N}{2} \rfloor} \frac{1}{h} \le 2 \log \left\lfloor \frac{N}{2} \right\rfloor + 2 \le 4 \log N.$$

The second to last inequality is a well-known estimate for partial sums of the harmonic series.

Now we have

$$\begin{aligned} &\frac{c_{\boldsymbol{\gamma}}}{N} \prod_{j=1}^{s} \left(\beta_j + \left(1 + 2b^{\min\{w_j, m\}}\right) \gamma_j S_N \right) \\ &\le \frac{c_{\boldsymbol{\gamma}}}{N} \prod_{j=1}^{s} \left(1 + \gamma_j \left(1 + 4\left(1 + 2b^{w_j}\right) \log N\right)\right) \\ &\le \frac{c_{\boldsymbol{\gamma}}}{N} \prod_{j=1}^{s} (1 + 13 \gamma_j b^{w_j} \log N). \end{aligned}$$

Define

$$\sigma_d := 13 \sum_{j=d+1}^{\infty} \gamma_j b^{w_j} \quad \text{for} \quad d \ge 0.$$

From [4, p. 222] or [6, Lemma 3] we know that

$$\prod_{j=1}^{s} (1 + 13 \gamma_j b^{w_j} \log N) \le \left(1 + \sigma_d^{-1}\right)^d N^{(\sigma_0+1)\sigma_d}.$$

For $0 < \delta < 1$ choose d large enough such that $\sigma_d \le \frac{\delta}{\sigma_0+1}$. Then

$$\prod_{j=1}^{s} (1 + 13 \gamma_j b^{w_j} \log N) \le \tilde{c}_{\boldsymbol{\gamma},\delta} N^{\delta},$$

where $\tilde{c}_{\boldsymbol{\gamma},\delta}$ is independent of s and N. Thus we have

$$R^s_{N,\boldsymbol{\gamma}}(\mathbf{z}) \le c_{\boldsymbol{\gamma},\delta} N^{\delta - 1},$$

with $c_{\boldsymbol{\gamma},\delta} = c_{\boldsymbol{\gamma}} \cdot \tilde{c}_{\boldsymbol{\gamma},\delta}$ independent of s and N. We obtain $c_{\boldsymbol{\gamma},\delta} N^{\delta-1} \le \varepsilon$ and thus

$$D^*_{N,\boldsymbol{\gamma}}(\mathbf{z}) \le \varepsilon \quad \text{if} \quad N \ge (c_{\boldsymbol{\gamma},\delta} \varepsilon^{-1})^{\frac{1}{1-\delta}}.$$

Hence, if the $\gamma_j b^{w_j}$'s are summable, we always achieve strong polynomial tractability.

Remark 2. Whether the conditions on the γ_j's and w_j's can be mitigated while at least polynomial tractability still holds remains an unresolved problem.

Acknowledgements. The authors would like to thank to Josef Dick, Peter Kritzer and Friedrich Pillichshammer for their comments and suggestions.

REFERENCES

[1] ARONSZAJN, N.: *Theory of reproducing kernels*, Trans. Amer. Math. Soc. **68** (1950), 337–404.
[2] DICK, J.—KRITZER, V—LEOBACHER, G.—PILLICHSHAMMER, F.: *A reduced fast component-by-component construction of lattice points for integration in weighted spaces with fast decreasing weights*, J. Comput. Appl. Math. **276** (2015), 1–15.
[3] DICK, J.—KUO, F. Y.—SLOAN, I. H.: *High-dimensional integration: the quasi-Monte Carlo way*, Acta Numer. **22** (2013), 133–288.
[4] DICK, J.—PILLICHSHAMMER, F.: *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press Cambridge, 2010.
[5] HICKERNELL, F. J.: *A generalized discrepancy and quadrature error bound*, Math. Comp. **67** (1998), 299–322.
[6] HICKERNELL, F. J.—NIEDERREITER, H.: *The existence of good extensible rank-1 lattices*, J. Complexity **19** (2003), 286–300.
[7] HLAWKA, E.: *Über die Diskrepanz mehrdimensionaler Folgen mod. 1*, Math. Z. **77** (1961), 273–284. (In German)
[8] HLAWKA, E.: *Zur angenherten Berechnung mehrfacher Integrale*, Monatsh. Math. **66** (1961), 140–151. (In German)
[9] JOE, S.: *Construction of good rank-1 lattice rules based on the weighted star discrepancy*. In: *Monte Carlo and Quasi-Monte Carlo Methods 2004* (H. Niederreiter and D. Talay, eds.), Springer, Berlin, 2006, pp. 181–196.
[10] KOROBOV, N. M.: *Approximate evaluation of repeated intergrals*, Dokl. Akad. Nauk SSSR **132** (1960), 1009–1012. (In Russian)
[11] LEMIEUX, C.: *Monte Carlo and Quasi-Monte Carlo Sampling, Springer Series in Statistics*. Springer, New York, 2009.
[12] LEOBACHER, G.—PILLICHSHAMMER, F.: *Introduction to Quasi-Monte Carlo Integration and Applications*. Compact Textbooks in Mathematics, Birkhuser, Cham, 2014.
[13] NIEDERREITER, H.: *Existence of good lattice points in the sense of Hlawka*, Monatsh. Math. **86** (1978), 203–219.
[14] NIEDERREITER, H.: *Random Number Generation and Quasi-Monte Carlo Methods*. In: *CBMS-NSF Regional Conference Series in Applied Mathematics Vol. 63*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1992.
[15] NOVAK, E.—WOŹNIAKOWSKI, H.: *Tractability of Multivariate Problems Vol. 1: Linear Information*. EMS, Zürich, 2008.
[16] NOVAK, E.—WOŹNIAKOWSKI, H.: *Tractability of Multivariate Problems Vol. 2: Standard Information for Functionals*. EMS, Zürich, 2010.
[17] NOVAK, E.—WOŹNIAKOWSKI, H.: *Tractability of Multivariate Problems Vol. 3: Standard Information for Operators*. EMS, Zürich, 2012.
[18] NUYENS, D.—COOLS, R.: *Fast component-by-component construction of rank-1 lattice rules with a non-prime number of points*, J. Complexity **22** (2006), 4–28.

[19] NUYENS, D.—COOLS, R.: *Fast algorithms for component-by-component construcion of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces*, Math. Comp. **75** (2006), 903–920.

[20] SINESCU, V.—JOE, S.: *Good lattice rules with a composite number of points based on the product weighted star discrepancy.* In: *Monte Carlo and Quasi-Monte Carlo Methods 2006* (A. Keller, S. Heinrich and H. Niederreiter, eds.), Springer, Berlin (2008), pp. 645–658.

[21] SINESCU, V.—JOE, S.: *Good lattice rules based on the general weighted star discrepancy*, Math. Comp. **76** (2007), 989–1004.

[22] SLOAN, I. H.—JOE, S.: *Lattice Methods for Multiple Integration.* Oxford Science Publications,
The Clarendon Press, Oxford University Press, New York, 1994.

[23] SLOAN, I. H.—WOŹNIAKOWSKI, H.: *When are quasi-Monte Carlo algorithms efficient for high-dimensional integrals?* J. Complexity **14** (1998), 1–33.

[24] ZAREMBA, S. K.: *Some applications of multidimensional integration by parts*, Ann. Polon. Math. **21** (1968), 85–96.

Received January 29, 2015
Accepted April 2, 2015

R. Kritzinger, H. Laimer
Department of Financial Mathematics and Applied Number Theory
Johannes Kepler University Linz
Altenbergerstr. 69
4040 Linz
AUSTRIA
E-mail: ralph.kritzinger@jku.at
helene.laimer@jku.at

Uniform Distribution Theory **10** (2015), no.2, 49–66

uniform distribution theory

COMPONENT-BY-COMPONENT CONSTRUCTION OF SHIFTED HALTON SEQUENCES

Peter Kritzer—Friedrich Pillichshammer

Dedicated to Professor Harald Niederreiter on the occasion of his 70th birthday

ABSTRACT. We study quasi-Monte Carlo integration in a weighted anchored Sobolev space. As the underlying integration nodes we consider Halton sequences in prime bases $\boldsymbol{p} = (p_1, \ldots, p_s)$ which are shifted with a $\boldsymbol{p}$-adic shift based on $\boldsymbol{p}$--adic arithmetic. The error is studied in the worst-case setting. In a recent paper, Hellekalek together with the authors of this article proved optimal error bounds in the root mean square sense, where the mean was extended over the uncountable set of all possible $\boldsymbol{p}$-adic shifts. Here we show that candidates for good shifts can in fact be chosen from a finite set and can be found by a component-by-component algorithm.

Communicated by Arne Winterhof

1. Introduction

We study the problem of approximating the value of the integral $I_s(f) := \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ of functions f belonging to a reproducing kernel Hilbert space $\mathcal{H}(K)$ of functions $f : [0,1]^s \to \mathbb{R}$. One way of numerically approximating $I_s(f)$ is to employ a quasi-Monte Carlo (QMC) rule,

$$Q_{N,s}(f) := \frac{1}{N} \sum_{n=0}^{N-1} f(\boldsymbol{x}_n),$$

where $P_{N,s} = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}\}$ is a set of N deterministically chosen points in $[0,1)^s$. It is well known (see, e. g., [3, 4, 10, 12, 15]) that point sets which are

2010 Mathematics Subject Classification: 65D30, 65C05, 11K38, 11K45.
Keywords: Quasi-Monte Carlo integration, shifted Halton sequences, worst-case error.
The authors are supported by the Austrian Science Fund (FWF): Projects F5506-N26 (Kritzer) and F5509-N26 (Pillichshammer), respectively, which are part of the Special Research Program "Quasi-Monte Carlo Methods: Theory and Applications".

in some way evenly distributed in the unit cube yield a low integration error when applying the corresponding QMC rules for approximating $I_s(f)$.

We study the error of QMC rules in the worst-case setting. The worst-case error of an algorithm $Q_{N,s}$ based on nodes $P_{N,s}$ is defined as the worst integration error over the unit ball of $\mathcal{H}(K)$, i. e.,

$$e_{N,s}(P_{N,s}, K) = \sup_{\substack{f \in \mathcal{H}(K) \\ \|f\|_K \leq 1}} |I_s(f) - Q_{N,s}(f)|.$$

An essential question in the theory of QMC methods is how the sample nodes $P_{N,s}$ of a QMC rule $Q_{N,s}$ should be chosen.

Shifted Halton sequences. In this paper we focus on a special kind of point sequences underlying a QMC rule, namely Halton sequences (cf. [5]) whose definition is based on the radical inverse function. Let $p \geq 2$ be an integer, $\mathbb{N} = \{1, 2, 3, \ldots\}$, and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For $n \in \mathbb{N}_0$, let $n = n_0 + n_1 p + n_2 p^2 + \cdots$ be the base p expansion of n (which is of course finite) with digits $n_i \in \{0, 1, \ldots \ldots, p-1\}$ for $i \geq 0$. The radical inverse function $\phi_p : \mathbb{N}_0 \to [0, 1)$ in base p is defined by

$$\phi_p(n) := \sum_{r=0}^{\infty} \frac{n_r}{p^{r+1}}.$$

Halton sequences can be defined for any dimension $s \in \mathbb{N}$. Let $p_1, \ldots, p_s \geq 2$ be s integers, and let $\boldsymbol{p} = (p_1, \ldots, p_s)$. Then the s-dimensional Halton sequence $H_{\boldsymbol{p}}$ in bases $p_1, \ldots, p_s$ is defined to be the sequence $H_{\boldsymbol{p}} = (\boldsymbol{x}_n)_{n \geq 0} \subseteq [0, 1)^s$, where

$$\boldsymbol{x}_n = \big(\phi_{p_1}(n), \phi_{p_2}(n), \ldots, \phi_{p_s}(n)\big), \quad \text{for } n \in \mathbb{N}_0.$$

It is well known (see, e.g., [3, 12]) that Halton sequences have good distribution properties if and only if the bases $p_1, \ldots, p_s$ are mutually relatively prime, and for the sake of simplicity we assume throughout the rest of the paper that $\boldsymbol{p} = (p_1, \ldots, p_s)$ consists of s mutually different prime numbers.

We also introduce a method of randomizing the elements of the Halton sequence which is referred to as a $\boldsymbol{p}$-adic shift. This special case of randomization is based on arithmetic over the p-adic numbers and is perfectly suited for Halton sequences $H_{\boldsymbol{p}}$.

Let p be a prime number. We define the set of p-adic numbers as the set of formal sums

$$\mathbb{Z}_p = \left\{ z = \sum_{r=0}^{\infty} z_r p^r \ : \ z_r \in \{0, 1, \ldots, p-1\} \text{ for all } r \in \mathbb{N}_0 \right\}.$$

Clearly $\mathbb{N}_0 \subseteq \mathbb{Z}_p$. For two nonnegative integers $y, z \in \mathbb{N}_0 \subseteq \mathbb{Z}_p$, the sum $y+z \in \mathbb{Z}_p$ is defined as the usual sum of integers. The addition can be extended to all p-adic numbers. The set $\mathbb{Z}_p$ with this addition, which we denote by $+_{\mathbb{Z}_p}$, then forms an abelian group.

As an extension of the radical inverse function defined above, we define the so-called Monna map

$$\phi_p : \mathbb{Z}_p \to [0,1) \quad \text{by} \quad \phi_p(z) := \sum_{r=0}^{\infty} \frac{z_r}{p^{r+1}} \pmod 1$$

whose restriction to $\mathbb{N}_0$ is exactly the radical inverse function in base p. In order to keep the used notation at a minimum we denote both, the Monna map and the radical inverse function, by ϕ_p. We also define the inverse

$$\phi_p^+ : [0,1) \to \mathbb{Z}_p \quad \text{by} \quad \phi_p^+ \left(\sum_{r=0}^{\infty} \frac{x_r}{p^{r+1}} \right) := \sum_{r=0}^{\infty} x_r p^r,$$

where we always use the finite p-adic representation for p-adic rationals in $[0,1)$. By a p-adic rational, we understand a number in $[0,1)$ that can be represented by a finite p-adic expansion.

For a prime number p and for $x \in [0,1)$ we consider the following p-adic shifts:

- **p-adic shift:** for $\sigma \in [0,1)$, we define $x \oplus_p \sigma \in [0,1)$ to be
$$x \oplus_p \sigma = \phi_p\big(\phi_p^+(x) +_{\mathbb{Z}_p} \phi_p^+(\sigma)\big).$$
- **simplified p-adic shift:** for $m \in \mathbb{N}$ and $\sigma \in [0,1)$, we write $x \oplus_{p,m}^{\mathrm{smp}} \sigma$ to be the truncation of $x \oplus_p \sigma$ to the m most significant digits, i.e., if $\phi_p^+(x) +_{\mathbb{Z}_p} \phi_p^+ \sigma = \sum_{r=0}^{\infty} y_r p^r \in \mathbb{Z}_p$, then
$$x \oplus_{p,m}^{\mathrm{smp}} \sigma = \phi_p \left(\sum_{r=0}^{m-1} y_r p^r \right).$$
- **mid-simplified p-adic shift:** for $m \in \mathbb{N}$ and $\sigma \in [0,1)$, we write
$$x \oplus_{p,m}^{\mathrm{mid}} \sigma = (x \oplus_{p,m}^{\mathrm{smp}} \sigma) + \frac{1}{2p^m}.$$

If the choice of m is clear from the context, we may often omit m in the notation $\oplus_{p,m}^{\mathrm{smp}}$ and $\oplus_{p,m}^{\mathrm{mid}}$ and write $\oplus_p^{\mathrm{smp}}$ and $\oplus_p^{\mathrm{mid}}$ instead.

In the s-variate case, for given bases $\boldsymbol{p} = (p_1, \ldots, p_s)$, a given point $\boldsymbol{x} = (x_1, \ldots, x_s) \in [0,1)^s$, and given $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_s) \in [0,1)^s$ and $\boldsymbol{m} = (m_1, \ldots, m_s) \in \mathbb{N}^s$, the above shifts are defined component-wise and we write $\boldsymbol{x} \oplus_{\boldsymbol{p}} \boldsymbol{\sigma} \in [0,1)^s$, $\boldsymbol{x} \oplus_{\boldsymbol{p},\boldsymbol{m}}^{\mathrm{smp}} \boldsymbol{\sigma}$ and $\boldsymbol{x} \oplus_{\boldsymbol{p},\boldsymbol{m}}^{\mathrm{mid}} \boldsymbol{\sigma}$, respectively.

For a point set $Y = \{\boldsymbol{y}_n \,:\, n = 0, \ldots, N-1\}$ we write

$$Y \oplus \boldsymbol{\sigma} := \{\boldsymbol{y}_n \oplus \boldsymbol{\sigma} \,:\, n = 0, \ldots, N-1\} \quad \text{where} \quad \oplus \text{ is either } \oplus_{\boldsymbol{p}}, \oplus_{\boldsymbol{p},\boldsymbol{m}}^{\mathrm{smp}}, \text{ or } \oplus_{\boldsymbol{p},\boldsymbol{m}}^{\mathrm{mid}}.$$

A weighted Sobolev space. In this paper, we are going to consider the problem of numerical integration of functions f that belong to a weighted anchored Sobolev space. Before we give the definition we introduce some notation which we require for the following: assume that $\boldsymbol{\gamma} = (\gamma_j)_{j=1}^{\infty}$ is a non-increasing sequence of positive weights, where $1 \geq \gamma_1 \geq \gamma_2 \geq \cdots$. These weights are used in order to model the influence of the different variables of the integrands, an idea which was introduced by Sloan and Woźniakowski [17]. For $s \in \mathbb{N}$ let $[s] := \{1, \ldots, s\}$. For $\mathfrak{u} \subseteq [s]$, $\boldsymbol{x}_{\mathfrak{u}}$ denotes the projection of $\boldsymbol{x} \in [0,1]^s$ onto $[0,1]^{|\mathfrak{u}|}$ consisting of the components whose indices are contained in $\mathfrak{u}$. Furthermore we write $(\boldsymbol{x}_{\mathfrak{u}}, \mathbf{1}) \in [0,1]^s$ for the point where those components of $\boldsymbol{x}$ whose indices are not in $\mathfrak{u}$ are replaced by 1.

We consider a weighted anchored Sobolev space $\mathcal{H}(K_{s,\boldsymbol{\gamma}})$ with anchor $\mathbf{1} = (1, 1, \ldots, 1)$ consisting of functions on $[0,1]^s$ whose first mixed partial derivatives are square integrable. This space is a reproducing kernel Hilbert space with kernel function

$$K_{s,\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{j=1}^{s} (1 + \gamma_j \min\bigl(1 - x_j, 1 - y_j\bigr)) \quad \text{for } \boldsymbol{x}, \boldsymbol{y} \in [0,1]^s, \tag{1}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_s)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_s)$. The inner product is given by

$$\langle f, g \rangle_{K_{s,\boldsymbol{\gamma}}} = \sum_{\mathfrak{u} \subseteq [s]} \gamma_{\mathfrak{u}}^{-1} \int_{[0,1]^{|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{x}_{\mathfrak{u}}} f(\boldsymbol{x}_{\mathfrak{u}}, \mathbf{1}) \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{x}_{\mathfrak{u}}} g(\boldsymbol{x}_{\mathfrak{u}}, \mathbf{1}) \, \mathrm{d}\boldsymbol{x}_{\mathfrak{u}}.$$

Here $\gamma_{\mathfrak{u}} = \prod_{j \in \mathfrak{u}} \gamma_j$; in particular $\gamma_{\emptyset} = 1$. Furthermore, we denote by $\frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{x}_{\mathfrak{u}}} h$ the derivative of a function h with respect to the x_j with $j \in \mathfrak{u}$. The norm in $\mathcal{H}(K_{s,\boldsymbol{\gamma}})$ is given by $\|f\|_{K_{s,\boldsymbol{\gamma}}} = \sqrt{\langle f, f \rangle_{K_{s,\boldsymbol{\gamma}}}}$. The Sobolev space $\mathcal{H}(K_{s,\boldsymbol{\gamma}})$ has been studied frequently in the literature (see, among many references, e. g., [1, 2, 7, 9, 11, 13, 17, 18]).

It is well known that the squared worst-case integration error in a reproducing kernel Hilbert space can be expressed in terms of the kernel function. In the particular case of the kernel $K_{s,\boldsymbol{\gamma}}$, it is easily derived with the help of [3, Proposition 2.11] that for $P_{N,s} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\}$ in $[0,1)^s$, where $\boldsymbol{x}_n = (x_{n,1}, \ldots, x_{n,s})$ for $n = 0, 1, \ldots, N-1$, we have

$$\begin{aligned} e^2_{N,s}(P_{N,s}, K_{s,\boldsymbol{\gamma}}) \;=\; & \prod_{i=1}^{s} \Bigl(1 + \frac{\gamma_i}{3}\Bigr) - \frac{2}{N} \sum_{n=0}^{N-1} \prod_{i=1}^{s} \Bigl(1 + \frac{\gamma_i}{2}(1 - x_{n,i}^2)\Bigr) \\ & + \frac{1}{N^2} \sum_{n,h=0}^{N-1} \prod_{i=1}^{s} \bigl(1 + \gamma_i \min(1 - x_{n,i}, 1 - x_{h,i})\bigr). \end{aligned} \tag{2}$$

Hence the worst-case error can be computed at a cost of $O(sN^2)$ arithmetic operations.

In [7] the authors studied the root mean square worst-case error in $\mathcal{H}(K_{s,\boldsymbol{\gamma}})$ of the $\boldsymbol{p}$-adically shifted Halton sequence extended over all $\boldsymbol{p}$-adic shifts, i.e.,

$$\widehat{e}_{N,s}(H_{\boldsymbol{p}}, K_{s,\boldsymbol{\gamma}}) := \sqrt{\mathbb{E}_{\boldsymbol{\sigma}}[e^2_{N,s}(H_{\boldsymbol{p}} \oplus_{\boldsymbol{p}} \boldsymbol{\sigma}, K_{s,\boldsymbol{\gamma}})]}.$$

We remark that there are some relations of $\widehat{e}_{N,s}$ to other figures of merit like the weighted L_2-discrepancy or the worst-case error in a certain reproducing kernel Hilbert space which is based on the so-called $\boldsymbol{p}$-adic function system (see [7] for more information). The latter one is a generalization of the $\boldsymbol{p}$-adic diaphony which was introduced by Hellekalek [6] (see also [14]).

The following result is the main result of [7].

Theorem 1 ([7, Theorem 1]). *Let $N \geq 2$. We have*

$$[\widehat{e}_{N,s}(H_{\boldsymbol{p}}, K_{s,\boldsymbol{\gamma}})]^2 \leq \frac{1}{N^2}\left[\prod_{j=1}^{s}\left(1+\gamma_j(\log N)\frac{p_j^2}{\log p_j}\right) + \prod_{j=1}^{s}\left(1+\frac{\gamma_j}{2}\right)\prod_{j=1}^{s}\left(1+\frac{\gamma_j p_j}{6}\right)\right]. \quad (3)$$

In particular, if $\sum_{j=1}^{\infty}\gamma_j\frac{p_j^2}{\log p_j} < \infty$, then for any $\delta > 0$ we have

$$\widehat{e}_{N,s}(H_{\boldsymbol{p}}, K_{s,\boldsymbol{\gamma}}) \ll_{\delta,\boldsymbol{\gamma},\boldsymbol{p}} \frac{1}{N^{1-\delta}},$$

where the implied constant is independent of the dimension s.

The bound (3) is, up to log-factors, optimal. For a further discussion of the result, especially with respect to the dependence on the dimension s we refer to [7]. Theorem 1 can also be interpreted in the "deterministic" sense that for every fixed $N \geq 2$ there exists a $\boldsymbol{p}$-adic shift $\boldsymbol{\sigma} \in [0,1)^s$ such that the squared worst-case error of the initial N elements of the corresponding $\boldsymbol{p}$-adically shifted Halton sequence satisfies the bound (3). The problem with this interpretation is that the $\boldsymbol{p}$-adic shift has to be chosen from an uncountable set, namely the s-dimensional unit cube. This is a big drawback if one wants to effectively find good $\boldsymbol{p}$-adic shifts.

It is the aim of this short paper to show that it suffices to choose the $\boldsymbol{p}$-adic shifts, which yield an upper bound of the form (3), from a finite set. This set of possible candidates has size N^s which is of course huge already for moderately large s or N. However we also show that in principle good shifts can be found by a component-by-component (CBC) algorithm. This idea is borrowed from the construction of good lattice point sets which goes back to Korobov [8] and

to Sloan and Reztsov [16], and which is nowadays used in a multitude of papers. With this "adaptive search" the search space is only of a size of order $O(sN)$.

The rest of the paper is structured as follows: In Section 2 we prove some auxiliary results. The CBC construction of $\boldsymbol{p}$-adic shifts as well as the statement and proof of the main results of this paper are presented in Section 3.

2. Auxiliary results

We use the following notation: for $p \in \mathbb{N}$ and $m \in \mathbb{N}_0$ let

$$\mathbb{Q}(p^m) := \{ap^{-m} \,:\, a = 0, 1, \ldots, p^m - 1\}.$$

We now show the following lemma.

Lemma 1. *Let $H_{p,N}$ be the point set consisting of the first N elements of H_p and let $m \in \mathbb{N}$ be minimal such that $N < p^m$. Furthermore, let $\sigma_m \in \mathbb{Q}(p^m)$. Then it is true that*

$$e^2_{N,1}(H_{p,N} \oplus_p^{\mathrm{mid}} \sigma_m, K_{1,\gamma_1}) \le p^m \int_0^{p^{-m}} e^2_{N,1}\big(H_{p,N} \oplus_p (\sigma_m + \delta), K_{1,\gamma_1}\big)\,\mathrm{d}\delta.$$

Proof. Let $H_{p,N} = \{h_0, h_1, \ldots, h_{N-1}\}$. From (2) we obtain

$$\begin{aligned}
&p^m \int_0^{p^{-m}} e^2_{N,1}\big(H_{p,N} \oplus_p (\sigma_m + \delta), K_{1,\gamma_1}\big)\,\mathrm{d}\delta \\
&= \Big(1 + \frac{\gamma_1}{3}\Big) - \frac{2}{N} \sum_{n=0}^{N-1} p^m \int_0^{p^{-m}} \Big(1 + \frac{\gamma_1}{2}\Big(1 - \big(h_n \oplus_p (\sigma_m + \delta)\big)^2\Big)\Big)\,\mathrm{d}\delta \\
&\quad + \frac{1}{N^2} \sum_{n=0}^{N-1} p^m \int_0^{p^{-m}} \Big(1 + \gamma_1\Big(1 - \big(h_n \oplus_p (\sigma_m + \delta)\big)\Big)\Big)\,\mathrm{d}\delta \\
&\quad + \frac{1}{N^2} \sum_{\substack{n,k=0 \\ n \ne k}}^{N-1} p^m \int_0^{p^{-m}} \Big(1 + \gamma_1 \min\big\{1 - \big(h_n \oplus_p (\sigma_m + \delta)\big), \\
&\qquad\qquad 1 - \big(h_k \oplus_p (\sigma_m + \delta)\big)\big\}\Big)\,\mathrm{d}\delta.
\end{aligned}$$

For given $n \in \{0, 1, \ldots, N-1\}$, let us now analyze the quantity

$$h_n \oplus_p (\sigma_m + \delta) = \phi_p\big(\phi_p^+(h_n) +_{\mathbb{Z}_p} \phi_p^+(\sigma_m + \delta)\big).$$

The base p expansion of h_n is of the form $h_n = \sum_{r=1}^{m} \frac{h_n^{(r)}}{p^r}$, since $N < p^m$. Furthermore, the base p expansions of σ_m and δ, respectively, are of the form

$$\sigma_m = \sum_{r=0}^{m-1} \frac{\sigma^{(r)}}{p^{r+1}} \quad \text{and} \quad \delta = \sum_{r=m}^{\infty} \frac{\delta^{(r)}}{p^{r+1}},$$

due to the assumptions on σ_m and δ. Consequently,

$$\phi_p^+(h_n) = \sum_{r=0}^{m-1} h_n^{(r)} p^r$$

and

$$\phi_p^+(\sigma_m + \delta) = \phi_p^+(\sigma_m) +_{\mathbb{Z}_p} \phi_p^+(\delta) = \sum_{r=0}^{m-1} \sigma^{(r)} p^r +_{\mathbb{Z}_p} \sum_{r=m}^{\infty} \delta^{(r)} p^r.$$

Let

$$\phi_p^+(h_n) +_{\mathbb{Z}_p} \phi_p^+(\sigma_m) = \sum_{r=0}^{m} y_r p^r$$

with $y_r \in \{0, 1, \ldots, p-1\}$. Then we obtain

$$\phi_p^+(h_n) +_{\mathbb{Z}_p} \phi_p^+(\sigma_m + \delta) = \sum_{r=0}^{m-1} y_r p^r +_{\mathbb{Z}_p} y_m p^m +_{\mathbb{Z}_p} \phi_p^+(\delta).$$

Note that $\sum_{r=0}^{m-1} y_r p^r$ is the truncation of the p-adic sum $\phi_p^+(h_n) +_{\mathbb{Z}_p} \phi_p^+(\sigma_m)$ to the first m digits. Hence

$$\phi_p \left(\sum_{r=0}^{m-1} y_r p^r \right) = h_n \oplus_p^{\mathrm{smp}} \sigma_m.$$

For short we write

$$\xi(h_n, \sigma_m) := \phi_p(y_{m+1} p^m).$$

Note that $\phi_p^+\big(\xi(h_n, \sigma_m)\big) = y_{m+1} p^m$. Hence we can write

$$h_n \oplus_p (\sigma_m + \delta) = \phi_p\big(\phi_p^+(h_n) +_{\mathbb{Z}_p} \phi_p^+(\sigma_m + \delta)\big) = (h_n \oplus_p^{\mathrm{smp}} \sigma_m) + \big(\xi(h_n, \sigma_m) \oplus_p \delta\big).$$

From this we obtain

$$\begin{aligned} & p^m \int_0^{p^{-m}} 1 + \frac{\gamma_1}{2} \Big(1 - \big(h_n \oplus_p (\sigma_m + \delta) \big)^2 \Big) \, \mathrm{d}\delta \\ = \;& p^m \int_0^{p^{-m}} 1 + \frac{\gamma_1}{2} \left(1 - \Big((h_n \oplus_p^{\mathrm{smp}} \sigma_m) + \big(\xi(h_n, \sigma_m) \oplus_p \delta \big) \Big)^2 \right) \mathrm{d}\delta. \end{aligned}$$

We now use [7, Lemma 3], which states that for any $f \in L_2([0,1])$ and any $y \in [0,1)$, we have

$$\int_0^1 f(x)\,\mathrm{d}x = \int_0^1 f(x \oplus_p y)\,\mathrm{d}x. \tag{4}$$

This yields

$$\begin{aligned} & p^m \int_0^{p^{-m}} 1 + \frac{\gamma_1}{2}\Big(1 - \big(h_n \oplus_p (\sigma_m + \delta)\big)^2\Big)\,\mathrm{d}\delta \\ &= \; p^m \int_0^{p^{-m}} 1 + \frac{\gamma_1}{2}\Big(1 - \big((h_n \oplus_p^{\mathrm{smp}} \sigma_m) + \delta\big)^2\Big)\,\mathrm{d}\delta \\ &= \; 1 + \frac{\gamma_1}{2}\big(1 - (h_n \oplus_p^{\mathrm{smp}} \sigma_m)^2\big) - \frac{1}{p^m}\frac{\gamma_1}{2}(h_n \oplus_p^{\mathrm{smp}} \sigma_m) - \frac{1}{p^{2m}}\frac{\gamma_1}{6}. \end{aligned}$$

Furthermore, in a similar fashion,

$$\begin{aligned} & p^m \int_0^{p^{-m}} 1 + \gamma_1\Big(1 - \big(h_n \oplus_p (\sigma_m + \delta)\big)\Big)\,\mathrm{d}\delta \\ &= \; p^m \int_0^{p^{-m}} 1 + \gamma_1\Big(1 - \big((h_n \oplus_p^{\mathrm{smp}} \sigma_m) + \big(\xi(h_n, \sigma_m) \oplus_p \delta\big)\big)\Big)\,\mathrm{d}\delta \\ &= \; p^m \int_0^{p^{-m}} 1 + \gamma_1\Big(1 - \big((h_n \oplus_p^{\mathrm{smp}} \sigma_m) + \delta\big)\Big)\,\mathrm{d}\delta \\ &= \; -\frac{\gamma_1}{2}\frac{1}{p^m} + 1 + \gamma_1 - \gamma_1(h_n \oplus_p^{\mathrm{smp}} \sigma_m). \end{aligned}$$

Finally, let us deal with the expression

$$p^m \int_0^{p^{-m}} 1 + \gamma_1 \min\big\{1 - \big(h_n \oplus_p (\sigma_m + \delta)\big), 1 - \big(h_k \oplus_p (\sigma_m + \delta)\big)\big\}\,\mathrm{d}\delta \tag{5}$$

with $k \neq n$. Note that, as $k \neq n$, we cannot have $h_n \oplus_p (\sigma_m + \delta) = h_k \oplus_p (\sigma_m + \delta)$. Suppose that

$$h_n \oplus_p (\sigma_m + \delta) < h_k \oplus_p (\sigma_m + \delta). \tag{6}$$

Using the notation introduced above, we can rewrite (6) as

$$(h_n \oplus_p^{\mathrm{smp}} \sigma_m) + \big(\xi(h_n, \sigma_m) \oplus_p \delta\big) < (h_k \oplus_p^{\mathrm{smp}} \sigma_m) + \big(\xi(h_k, \sigma_m) \oplus_p \delta\big).$$

Again, since $k \neq n$, we cannot have

$$(h_n \oplus_p^{\mathrm{smp}} \sigma_m) = (h_k \oplus_p^{\mathrm{smp}} \sigma_m),$$

as this would also imply $\xi(h_n, \sigma_m) = \xi(h_k, \sigma_m)$, and so would yield a contradiction to (6). Furthermore, it cannot be the case that

$$(h_n \oplus_p^{\mathrm{smp}} \sigma_m) > (h_k \oplus_p^{\mathrm{smp}} \sigma_m),$$

since $\xi(h_n,\sigma_m),\xi(h_k,\sigma_m)\in[0,p^{-m})$, and so we would also end up with a contradiction to (6). Therefore, we see that (6) automatically implies

$$(h_n \oplus_p^{\mathrm{smp}} \sigma_m) < (h_k \oplus_p^{\mathrm{smp}} \sigma_m). \tag{7}$$

Suppose now, on the other hand, that (7) holds. Then, since $\xi(h_n,\sigma_m)$ and $\xi(h_k,\sigma_m)$ are in $[0,p^{-m})$, also (6) must hold. We have thus shown that (6) and (7) are equivalent.

Suppose now in the analysis of (5) that (6) holds, i. e.,

$$\begin{aligned}
&p^m\int_0^{p^{-m}} 1+\gamma_1\min\Big\{1-\big(h_n\oplus_p(\sigma_m+\delta)\big),1-\big(h_k\oplus_p(\sigma_m+\delta)\big)\Big\}\,\mathrm{d}\delta\\
&=p^m\int_0^{p^{-m}} 1+\gamma_1\Big(1-\big(h_k\oplus_p(\sigma_m+\delta)\big)\Big)\,\mathrm{d}\delta\\
&=p^m\int_0^{p^{-m}} 1+\gamma_1\Big(1-\Big((h_k\oplus_p^{\mathrm{smp}}\sigma_m)+\big(\xi(h_k,\sigma_m)\oplus_p\delta\big)\Big)\Big)\,\mathrm{d}\delta.
\end{aligned}$$

Using the equivalence between (6) and (7), and again (4), we see that the latter expression equals

$$\begin{aligned}
&p^m\int_0^{p^{-m}}\Big(1+\gamma_1\Big(\min\big\{1-(h_n\oplus_p^{\mathrm{smp}}\sigma_m),1-(h_k\oplus_p^{\mathrm{smp}}\sigma_m)\big\}-\big(\xi(h_k,\sigma_m)\oplus_p\delta\big)\Big)\Big)\,\mathrm{d}\delta\\
&=p^m\int_0^{p^{-m}} 1+\gamma_1\left(\min\left\{1-(h_n\oplus_p^{\mathrm{smp}}\sigma_m),1-(h_k\oplus_p^{\mathrm{smp}}\sigma_m)\right\}-\delta\right)\mathrm{d}\delta\\
&=-\frac{\gamma_1}{2}\frac{1}{p^m}+1+\gamma_1\min\left\{1-(h_n\oplus_p^{\mathrm{smp}}\sigma_m),1-(h_k\oplus_p^{\mathrm{smp}}\sigma_m)\right\}.
\end{aligned}$$

A similar argument holds if the converse of (6) holds.

Putting all of these observations together, we obtain

$$\begin{aligned}
p^m\int_0^{p^{-m}} e_{N,1}^2\big(H_{p,N}\oplus_p(\sigma_m+\delta),K_{1,\gamma_1}\big)\,\mathrm{d}\delta &= \Big(1+\frac{\gamma_1}{3}\Big)\\
&-\frac{2}{N}\sum_{n=0}^{N-1}\left(1+\frac{\gamma_1}{2}\left(1-(h_n\oplus_p^{\mathrm{smp}}\sigma_m)^2\right)-\frac{\gamma_1}{2}\frac{1}{p^m}(h_n\oplus_p^{\mathrm{smp}}\sigma_m)-\frac{1}{p^{2m}}\frac{\gamma_1}{6}\right)\\
&+\frac{1}{N^2}\sum_{n=0}^{N-1}\left(-\frac{\gamma_1}{2}\frac{1}{p^m}+1+\gamma_1-\gamma_1(h_n\oplus_p^{\mathrm{smp}}\sigma_m)\right)\\
&+\frac{1}{N^2}\sum_{\substack{n,k=0\\ n\neq k}}^{N-1}\left(-\frac{\gamma_1}{2}\frac{1}{p^m}+1+\gamma_1\min\left\{1-(h_n\oplus_p^{\mathrm{smp}}\sigma_m),1-(h_k\oplus_p^{\mathrm{smp}}\sigma_m)\right\}\right)
\end{aligned}$$

$$
\begin{aligned}
&\geq \left(1+\frac{\gamma_1}{3}\right) \\
&\quad - \frac{2}{N}\sum_{n=0}^{N-1}\left(1+\frac{\gamma_1}{2}\left(1-(h_n \oplus_p^{\mathrm{smp}} \sigma_m)^2\right) - \frac{\gamma_1}{2}\frac{1}{2p^m}2(h_n \oplus_p^{\mathrm{smp}} \sigma_m) - \frac{\gamma_1}{2}\frac{1}{4p^{2m}}\right) \\
&\quad + \frac{1}{N^2}\sum_{n=0}^{N-1}\left(1+\gamma_1\left(1-\left(h_n \oplus_p^{\mathrm{smp}} \sigma_m + \frac{1}{2p^m}\right)\right)\right) \\
&\quad + \frac{1}{N^2}\sum_{\substack{n,k=0\\ n\neq k}}^{N-1}\left(1+\gamma_1 \min\left\{1-\left(h_n \oplus_p^{\mathrm{smp}} \sigma_m + \frac{1}{2p^m}\right),\right.\right. \\
&\qquad\qquad \left.\left. 1-\left(h_k \oplus_p^{\mathrm{smp}} \sigma_m + \frac{1}{2p^m}\right)\right\}\right) \\
&= \left(1+\frac{\gamma_1}{3}\right) - \frac{2}{N}\sum_{n=0}^{N-1}\left(1+\frac{\gamma_1}{2}\left(1-\left(h_n \oplus_p^{\mathrm{mid}} \sigma_m\right)^2\right)\right) \\
&\quad + \frac{1}{N^2}\sum_{n,k=0}^{N-1}\left(1+\gamma_1 \min\left\{1-\left(h_n \oplus_p^{\mathrm{mid}} \sigma_m\right), 1-\left(h_k \oplus_p^{\mathrm{mid}} \sigma_m\right)\right\}\right) \\
&= e_{N,1}^2(H_{p,N} \oplus_p^{\mathrm{mid}} \sigma_m, K_{1,\gamma_1}).
\end{aligned}
$$

The result follows. □

For two point sets

$$X = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}\} \text{ in } [0,1)^{s_1} \quad \text{and} \quad Y = \{\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{N-1}\} \text{ in } [0,1)^{s_2}$$

we write (X,Y) to denote the point set consisting of the concatenated points

$$(\boldsymbol{x}_k, \boldsymbol{y}_k) = (x_{k,1}, \ldots, x_{k,s_1}, y_{k,1}, \ldots, y_{k,s_2}) \in [0,1)^{s_1+s_2} \quad \text{for } k = 0,1,\ldots,N-1.$$

Using this notation, we have the following result.

Lemma 2. *Let $P_{s,N}$ be a point set of N points in $[0,1)^s$. Let $H_{p,N}$ be as in Lemma 1 and let $m \in \mathbb{N}$ be minimal such that $N < p^m$. Furthermore, let $\sigma_m \in \mathbb{Q}(p^m)$. Then it is true that*

$$
\begin{aligned}
&e_{N,s+1}^2\left((P_{s,N}, H_{p,N} \oplus_p^{\mathrm{mid}} \sigma_m), K_{s+1,\boldsymbol{\gamma}}\right) \\
&\qquad\qquad \leq p^m \int_0^{p^{-m}} e_{N,s+1}^2\Big(\big(P_{s,N}, H_{p,N} \oplus_p (\sigma_m + \delta)\big), K_{s+1,\boldsymbol{\gamma}}\Big)\,\mathrm{d}\delta.
\end{aligned}
$$

P r o o f. The proof is similar to that of Lemma 1. □

3. The CBC construction

In this section, we analyze the following CBC construction of a mid-simplified $\boldsymbol{p}$-adic shift to obtain $\boldsymbol{p}$-adically shifted Halton sequences with a low integration error.

Throughout this section, let $s, N \in \mathbb{N}$ be given and let $\boldsymbol{p} = (p_1, \dots, p_s)$ with pairwise distinct prime components p_j. For $j \in [s]$ let $m_j \in \mathbb{N}$ be minimal such that $N < p_j^{m_j}$. Let $H_{\boldsymbol{p},N}$ be the point set consisting of the first N elements of $H_{\boldsymbol{p}}$. To stress the dependence of the worst-case error on the $\boldsymbol{p}$-adic shift we write in the following

$$e_{N,s}(\boldsymbol{\sigma}) := e_{N,s}(H_{\boldsymbol{p},N} \oplus_{\boldsymbol{p}}^{\mathrm{mid}} \boldsymbol{\sigma}, K_{s,\boldsymbol{\gamma}}) \quad \text{for } \boldsymbol{\sigma} \in \mathbb{Q}(p_1^{m_1}) \times \cdots \times \mathbb{Q}(p_s^{m_s}).$$

We propose the following algorithm.

Algorithm 1.

(1) Choose $\sigma_1 \in \mathbb{Q}(p_1^{m_1})$ to minimize $e_{N,1}^2(\sigma)$ as a function of σ.

(2) For $1 \le d \le s-1$, assume that $\sigma_1, \dots, \sigma_d$ have already been found. Choose

$$\sigma_{d+1} \in \mathbb{Q}(p_{d+1}^{m_{d+1}})$$

to minimize

$$e_{N,d+1}^2((\sigma_1, \dots, \sigma_d, \sigma)) \tag{8}$$

as a function of σ.

(3) If $d \le s-1$ increase d by 1 and go to Step 2, otherwise stop.

Remark 1. We remark that Algorithm 1 makes the main result in [7] much more explicit, as the algorithm only needs to check a countable number of possible candidates for the $\boldsymbol{p}$-adic shift. A slight drawback of our method is that the effective CBC construction of good $\boldsymbol{p}$-adic shifts has a cost of $O(s^2N^3)$ operations, which is still large. Using a probabilistic version of Algorithm 1 leads to a reduction of a factor N in the construction cost, see Algorithm 2. Further improvements with respect to the construction cost are a demanding problem for future research.

The following theorem states that Algorithm 1 yields $\boldsymbol{p}$-adically shifted Halton sequences with a low integration error. Note that the error bound is of the same order as the one in Theorem 1.

Theorem 2. *Let the notation be as above, and let $d \in [s]$. Assume that*

$$\boldsymbol{\sigma}_s = (\sigma_1, \ldots, \sigma_s)$$

has been constructed according to Algorithm 1. Then

$$e_{N,d}(\boldsymbol{\sigma}_d) \le \frac{1}{N} \left(\prod_{j=1}^{d} \left(1 + 2\gamma_j (\log N) \frac{p_j^2}{\log p_j} \right) + \prod_{j=1}^{d} (1+\gamma_j) \prod_{j=1}^{d} \left(1 + \frac{\gamma_j p_j}{6} \right) \right)^{1/2},$$

where $\boldsymbol{\sigma}_d := (\sigma_1, \ldots, \sigma_d)$.

P r o o f. We show the result by induction on d. For $d = 1$ we have

$$\begin{aligned}
&\int_0^1 e_{N,1}^2(H_{p_1,N} \oplus_{p_1} \sigma, K_{1,\gamma_1}) \, \mathrm{d}\sigma \\
&\quad = \frac{1}{p_1^{m_1}} \sum_{\ell=0}^{p_1^{m_1}-1} p_1^{m_1} \int_{\ell/p_1^{m_1}}^{(\ell+1)/p_1^{m_1}} e_{N,1}^2 \left(H_{p_1,N} \oplus_{p_1} \left(\frac{\ell}{p_1^{m_1}} + \delta \right), K_{1,\gamma_1} \right) \mathrm{d}\delta \\
&\quad \ge \frac{1}{p_1^{m_1}} \sum_{\ell=0}^{p_1^{m_1}-1} e_{N,1}^2 \left(\frac{\ell}{p_1^{m_1}} \right),
\end{aligned}$$

where we applied Lemma 1. Hence there exists a $\sigma_1' \in \mathbb{Q}(p_1^{m_1})$ such that

$$\begin{aligned}
e_{N,1}^2(\sigma_1') &\le \int_0^1 e_{N,1}^2(H_{p_1,N} \oplus_{p_1} \sigma, K_{1,\gamma_1}) \, \mathrm{d}\sigma \\
&\le \frac{1}{N^2} \left(1 + 2\gamma_1 (\log N) \frac{p_1^2}{\log p_1} + (1+\gamma_1) \left(1 + \frac{\gamma_1 p_1}{6} \right) \right),
\end{aligned}$$

where we used [7, Theorem 1] for the second inequality. Since σ_1 is chosen by Algorithm 1 to minimize $e_{N,1}^2(\sigma)$, it follows that the result holds for $d = 1$.

Suppose the result has already been shown for some fixed $d \in [s-1]$. Assume that $\boldsymbol{\sigma}_d = (\sigma_1, \ldots, \sigma_d)$ has been obtained by the CBC algorithm. Since σ_{d+1} is chosen in order to minimize the squared error (8), we have (where we write with some abuse of notation $(\boldsymbol{\sigma}_d, \sigma_{d+1}) := (\sigma_1, \ldots, \sigma_d, \sigma_{d+1})$)

$$e_{N,d+1}^2((\boldsymbol{\sigma}_d, \sigma_{d+1})) \le \frac{1}{p_{d+1}^{m_{d+1}}} \sum_{v=0}^{p_{d+1}^{m_{d+1}}-1} e_{N,d+1}^2 \left(\left(\boldsymbol{\sigma}_d, \frac{v}{p_{d+1}^{m_{d+1}}} \right) \right).$$

Using Lemma 2, we now see that, for any $v \in \{0, \dots, p_{d+1}^{m_{d+1}} - 1\}$,

$$\begin{aligned}
&\frac{1}{p_{d+1}^{m_{d+1}}} e_{N,d+1}^2 \left(\left(\boldsymbol{\sigma}_d, \frac{v}{p_{d+1}^{m_{d+1}}} \right) \right) \\
&\le \int e_{N,d+1}^2 \left(\left(H_{\boldsymbol{p}_d,N} \oplus_{\boldsymbol{p}}^{\mathrm{mid}} \boldsymbol{\sigma}_d, H_{p_{d+1},N} \oplus_{p_{d+1}} \left(\frac{v}{p_{d+1}^{m_{d+1}}} + \delta \right) \right), K_{d+1,\boldsymbol{\gamma}} \right) \mathrm{d}\delta,
\end{aligned}$$

where the integral is between 0 and $p_{d+1}^{-m_{d+1}}$ and where $\boldsymbol{p}_d := (p_1, \dots, p_d)$. Hence

$$e_{N,d+1}^2((\boldsymbol{\sigma}_d, \sigma_{d+1})) \le \int_0^1 e_{N,d+1}^2\big((H_{\boldsymbol{p}_d,N} \oplus_{\boldsymbol{p}}^{\mathrm{mid}} \boldsymbol{\sigma}_d, H_{p_{d+1},N} \oplus_{p_{d+1}} \sigma), K_{d+1,\boldsymbol{\gamma}}\big)\, \mathrm{d}\sigma.$$

We denote the points of $H_{\boldsymbol{p}_d,N} \oplus_{\boldsymbol{p}}^{\mathrm{mid}} \boldsymbol{\sigma}_d$ by $\boldsymbol{x}_n = (x_{n,1}, \dots, x_{n,d})$, and the points of $H_{p_{d+1},N}$ by h_n. Due to (2), we obtain

$$\begin{aligned}
\int_0^1 e_{N,d+1}^2\big((H_{\boldsymbol{p}_d,N} \oplus_{\boldsymbol{p}}^{\mathrm{mid}} \boldsymbol{\sigma}_d, H_{p_{d+1},N} \oplus_{p_{d+1}} \sigma), K_{d+1,\boldsymbol{\gamma}}\big)\, \mathrm{d}\sigma &= \prod_{j=1}^{d+1} \Big(1 + \frac{\gamma_j}{3}\Big) \\
&- \frac{2}{N} \sum_{n=0}^{N-1} \left[\prod_{j=1}^{d} \Big(1 + \frac{\gamma_j}{2}(1 - x_{n,j}^2)\Big) \right] \int_0^1 \Big(1 + \frac{\gamma_{d+1}}{2}\big(1 - (h_n \oplus_{p_{d+1}} \sigma)^2\big)\Big)\, \mathrm{d}\sigma \\
&+ \frac{1}{N^2} \sum_{n,k=0}^{N-1} \left[\prod_{j=1}^{d} (1 + \gamma_j \min\{1 - x_{n,j}, 1 - x_{k,j}\}) \right] \\
&\qquad \times \int_0^1 \big(1 + \gamma_{d+1} \min\{1 - (h_n \oplus_{p_{d+1}} \sigma), 1 - (h_k \oplus_{p_{d+1}} \sigma)\}\big)\, \mathrm{d}\sigma.
\end{aligned}$$

Let now

$$I_1 := \int_0^1 \Big(1 + \frac{\gamma_{d+1}}{2}\big(1 - (h_n \oplus_{p_{d+1}} \sigma)^2\big)\Big)\, \mathrm{d}\sigma,$$

and

$$I_2 := \int_0^1 \big(1 + \gamma_{d+1} \min\{1 - (h_n \oplus_{p_{d+1}} \sigma), 1 - (h_k \oplus_{p_{d+1}} \sigma)\}\big)\, \mathrm{d}\sigma.$$

Using (4), we obtain

$$I_1 = \int_0^1 \Big(1 + \frac{\gamma_{d+1}}{2}(1 - \sigma^2)\Big)\, \mathrm{d}\sigma = 1 + \frac{\gamma_{d+1}}{3}.$$

Let us now deal with I_2. Applying Proposition 2 in [7] yields that

$$I_2 = \sum_{\ell=0}^{\infty} r_{p_{d+1},\gamma_{d+1}}(\ell) \beta_\ell(h_n) \overline{\beta_\ell(h_k)},$$

where for $\ell = \ell_{a-1}p_{d+1}^{a-1} + \cdots + \ell_1 p_{d+1} + \ell_0$ with $\ell_{a-1} \neq 0$ we have

$$r_{p_{d+1},\gamma_{d+1}} = \begin{cases} 1 + \frac{\gamma_{d+1}}{3} & \text{if } \ell = 0, \\ \frac{\gamma_{d+1}}{2p_{d+1}^a}\left(\frac{1}{\sin^2(\ell_{a-1}\pi/p_{d+1})} - \frac{1}{3}\right) & \text{if } \ell \neq 0. \end{cases}$$

Altogether, we obtain

$$\begin{aligned} & e_{N,d+1}^2((\boldsymbol{\sigma}_d, \sigma_{d+1})) \\ & \leq \prod_{j=1}^{d+1}\left(1+\frac{\gamma_j}{3}\right) - \frac{2}{N}\sum_{n=0}^{N-1}\left[\prod_{j=1}^{d}\left(1+\frac{\gamma_j}{2}(1-x_{n,j}^2)\right)\right]\left(1+\frac{\gamma_{d+1}}{3}\right) \\ & \quad + \frac{1}{N^2}\sum_{n,k=0}^{N-1}\left[\prod_{j=1}^{d}(1+\gamma_j\min\{1-x_{n,j}, 1-x_{k,j}\})\right] \\ & \qquad\qquad \times \sum_{\ell=0}^{\infty} r_{p_{d+1},\gamma_{d+1}}(\ell)\beta_\ell(h_n)\overline{\beta_\ell(h_k)} \\ & = \left(1+\frac{\gamma_{d+1}}{3}\right)\left[\prod_{j=1}^{d}\left(1+\frac{\gamma_j}{3}\right) - \frac{2}{N}\sum_{n=0}^{N-1}\prod_{j=1}^{d}\left(1+\frac{\gamma_j}{2}(1-x_{n,j}^2)\right)\right. \\ & \qquad \left. + \frac{1}{N^2}\sum_{n,k=0}^{N-1}\prod_{j=1}^{d}(1+\gamma_j\min\{1-x_{n,j}, 1-x_{k,j}\})\right] \\ & \quad + \frac{1}{N^2}\sum_{n,k=0}^{N-1}\left(\prod_{j=1}^{d}(1+\gamma_j\min\{1-x_{n,j}, 1-x_{k,j}\})\right) \\ & \qquad\qquad \times \sum_{\ell=1}^{\infty} r_{p_{d+1},\gamma_{d+1}}(\ell)\beta_\ell(h_n)\overline{\beta_\ell(h_k)} \\ & = \left(1+\frac{\gamma_{d+1}}{3}\right) e_{N,d}^2(\boldsymbol{\sigma}_d) + T, \end{aligned} \tag{9}$$

where

$$T := \frac{1}{N^2}\sum_{n,k=0}^{N-1}\left(\prod_{j=1}^{d}(1+\gamma_j\min\{1-x_{n,j}, 1-x_{k,j}\})\right) \times \sum_{\ell=1}^{\infty} r_{p_{d+1},\gamma_{d+1}}(\ell)\beta_\ell(h_n)\overline{\beta_\ell(h_k)}.$$

Since $\min\{1-x_{n,j}, 1-x_{k,j}\} \leq 1$ we obviously have

$$T \leq \left(\prod_{j=1}^{d}(1+\gamma_j)\right)\sum_{\ell=1}^{\infty} r_{p_{d+1},\gamma_{d+1}}(\ell)\left|\frac{1}{N}\sum_{n=0}^{N-1}\beta_\ell(h_n)\right|^2. \tag{10}$$

From the proof of [7, Theorem 1], it can easily be derived that

$$\sum_{\ell=1}^{\infty} r_{p_{d+1},\gamma_{d+1}}(\ell) \left| \frac{1}{N} \sum_{n=0}^{N-1} \beta_\ell(h_n) \right|^2 \le \frac{1}{N^2} \frac{\gamma_{d+1} g p_{d+1}^2}{2} + \frac{\gamma_{d+1}}{6 p_{d+1}^g} \left(1 + \frac{\gamma_{d+1}}{2}\right),$$

for arbitrarily chosen $g \in \mathbb{N}_0$. By choosing $g = \lfloor 2 \log_{p_{d+1}} N \rfloor$ and inserting into (10), we arrive at

$$\begin{aligned} T &\le \frac{1}{N^2} \prod_{j=1}^{d} (1+\gamma_j) \left(\left(\gamma_{d+1} (\log N) \frac{p_{d+1}^2}{\log p_{d+1}} \right) + \frac{\gamma_{d+1} p_{d+1}}{6} \left(1 + \frac{\gamma_{d+1}}{2}\right) \right) \\ &\le \frac{1}{N^2} \left(\left(\gamma_{d+1} (\log N) \frac{p_{d+1}^2}{\log p_{d+1}} \right) \prod_{j=1}^{d} \left(1 + 2\gamma_j (\log N) \frac{p_j^2}{\log p_j} \right) \right. \\ &\qquad \left. + \frac{\gamma_{d+1} p_{d+1}}{6} \prod_{j=1}^{d+1} (1+\gamma_j) \prod_{j=1}^{d} \left(1 + \frac{\gamma_j p_j}{6}\right) \right). \end{aligned} \tag{11}$$

On the other hand, we have, using the induction assumption,

$$\begin{aligned} &\left(1 + \frac{\gamma_{d+1}}{3}\right) e_{N,d}^2(\boldsymbol{\sigma}_d) \\ &\quad \le \left(1 + \frac{\gamma_{d+1}}{3}\right) \frac{1}{N^2} \\ &\qquad \left(\prod_{j=1}^{d} \left(1 + 2\gamma_j (\log N) \frac{p_j^2}{\log p_j} \right) + \prod_{j=1}^{d} (1+\gamma_j) \prod_{j=1}^{d} \left(1 + \frac{\gamma_j p_j}{6}\right) \right) \\ &\quad \le \frac{1}{N^2} \left(\left(1 + \gamma_{d+1} (\log N) \frac{p_{d+1}^2}{\log p_{d+1}} \right) \prod_{j=1}^{d} \left(1 + 2\gamma_j (\log N) \frac{p_j^2}{\log p_j} \right) \right. \\ &\qquad \left. + \prod_{j=1}^{d+1} (1+\gamma_j) \prod_{j=1}^{d} \left(1 + \frac{\gamma_j p_j}{6}\right) \right). \end{aligned} \tag{12}$$

Combining equations (11) and (12), and inserting into (9), we obtain

$$\begin{aligned} &e_{N,d+1}^2((\boldsymbol{\sigma}_d, \sigma_{d+1})) \\ &\quad \le \frac{1}{N^2} \left(\prod_{j=1}^{d+1} \left(1 + 2\gamma_j (\log N) \frac{p_j^2}{\log p_j} \right) + \prod_{j=1}^{d+1} (1+\gamma_j) \prod_{j=1}^{d+1} \left(1 + \frac{\gamma_j p_j}{6}\right) \right). \end{aligned}$$

Taking the square root we obtain the result for $d+1$, and the theorem is shown. □

We also propose the following randomized algorithm.

Algorithm 2. Let $c \in \mathbb{N}$ such that $c \leq p_i^{m_i}$ for all $i \in [s]$.

(1) Randomly choose c shifts $\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_c \in \mathbb{Q}(p_1^{m_1})$, where $\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_c$ are uniformly i.i.d. Set $\sigma_1 = \widetilde{\sigma}_u$, where $u \in \{1, 2, \ldots, c\}$ is the value of w which minimizes $e_{N,1}^2(\widetilde{\sigma}_w)$ over $w \in \{1, 2, \ldots, c\}$.

(2) For $1 \leq d \leq s-1$, assume that $\sigma_1, \ldots, \sigma_d$ have already been found. Randomly choose c shifts $\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_c \in \mathbb{Q}(p_{d+1}^{m_{d+1}})$, where $\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_c$ are uniformly i.i.d. Set $\sigma_{d+1} = \widetilde{\sigma}_u$, where $u \in \{1, 2, \ldots, c\}$ is the value of w which minimizes

$$e_{N,d+1}^2((\sigma_1, \ldots, \sigma_d, \widetilde{\sigma}_w))$$

over $w \in \{1, 2, \ldots, c\}$.

(3) If $d \leq s-1$ increase d by 1 and go to Step 2, otherwise stop.

Remark 2. We remark that Algorithm 2 has a reduced construction cost of $O(cs^2N^2)$ operations. This means, if c is fixed and small compared to N we save a factor of N in the construction cost compared to Algorithm 1. This reduction is penalized by a slightly worse error estimate which now only holds with a certain probability. This is the essence of the following theorem.

Theorem 3. *Let $t \geq 1$. The probability that the vector $\boldsymbol{\sigma}_s = (\sigma_1, \ldots, \sigma_s)$ constructed by Algorithm 2 satisfies*

$$e_{N,d}(\boldsymbol{\sigma}_d) \leq \frac{t}{N} \left(\prod_{j=1}^{d} \left(1 + 2\gamma_j (\log N) \frac{p_j^2}{\log p_j} \right) + \prod_{j=1}^{d} (1+\gamma_j) \prod_{j=1}^{d} \left(1 + \frac{\gamma_j p_j}{6} \right) \right)^{1/2}$$

for all $d \in [s]$, where $\boldsymbol{\sigma}_d := (\sigma_1, \ldots, \sigma_d)$, is at least

$$\left(1 - \frac{1}{t^{2c}}\right)^s \geq 1 - \frac{s}{t^{2c}}.$$

Proof. For $d \in [s]$ let

$$L_{d,N} := \prod_{j=1}^{d} \left(1 + 2\gamma_j (\log N) \frac{p_j^2}{\log p_j} \right) + \prod_{j=1}^{d} (1+\gamma_j) \prod_{j=1}^{d} \left(1 + \frac{\gamma_j p_j}{6} \right).$$

According to the proof of Theorem 2 we have

$$\frac{1}{p_1^{m_1}} \sum_{\ell=0}^{p_1^{m_1}-1} e_{N,1}^2 \left(\frac{\ell}{p_1^{m_1}} \right) \leq \frac{1}{N^2} L_{1,N}.$$

Using Markov's inequality we obtain that for all $t \geq 1$ we have

$$\frac{1}{p_1^{m_1}} \left| \left\{ \sigma \in \mathbb{Q}(p_1^{m_1}) \ : \ e_{N,1}^2(\sigma) \leq \frac{t}{N^2} L_{1,N} \right\} \right| \geq 1 - \frac{1}{t},$$

which can be re-written as

$$\frac{1}{p_1^{m_1}} \left| \left\{ \sigma \in \mathbb{Q}(p_1^{m_1}) \ : \ e_{N,1}(\sigma) \ \leq \frac{t}{N} L_{1,N}^{1/2} \right\} \right| \geq 1 - \frac{1}{t^2}.$$

Hence the probability that at least one of $\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_c$ satisfies

$$e_{N,1}(\widetilde{\sigma}_w) \ \leq \ \frac{t}{N} \ L_{1,N}^{1/2}$$

is at least $1 - t^{-2c}$.

Furthermore it was shown in the proof of Theorem 2 that under the assumption

$$e_{n,d}^2(\boldsymbol{\sigma}_d) \leq \frac{1}{N^2} L_{d,N}$$

we have

$$\frac{1}{p_{d+1}^{m_{d+1}}} \sum_{v=0}^{p_{d+1}^{m_{d+1}}-1} e_{N,d+1}^2 \left(\left(\boldsymbol{\sigma}_d, \frac{v}{p_{d+1}^{m_{d+1}}} \right) \right) \leq \frac{1}{N^2} L_{d+1,N}.$$

Using Markov's inequality again we obtain for $t \geq 1$

$$\frac{1}{p_{d+1}^{m_{d+1}}} \left| \left\{ \sigma \in \mathbb{Q}(p_{d+1}^{m_{d+1}}) \ : \ e_{N,d+1}((\boldsymbol{\sigma}_d, \sigma)) \leq \frac{t}{N} L_{d+1,N}^{1/2} \right\} \right| \geq 1 - \frac{1}{t^2}.$$

Hence the probability that at least one of $\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_c$ satisfies

$$e_{N,d+1}((\boldsymbol{\sigma}_d, \widetilde{\sigma}_w)) \leq \frac{t}{N} L_{d+1,N}^{1/2}$$

is at least $1 - t^{-2c}$. Hence the result follows. □

REFERENCES

[1] DICK, J.—KUO, F. Y.—PILLICHSHAMMER, F.—SLOAN, I. H.: *Construction algorithms for polynomial lattice rules for multivariate integration*, Math. Comp. **74** (2005), 1895–1921.

[2] DICK, J.—PILLICHSHAMMER, F.: *Multivariate integration in weighted Hilbert spaces based on Walsh functions and weighted Sobolev spaces*, J. Complexity **21** (2005), 149–195.

[3] DICK, J.—PILLICHSHAMMER, F.: *Digital Nets and Sequences. Discrepancy Theory and Quasi-Monte Carlo Integration*, Cambridge University Press, 2010.

[4] DRMOTA, M.— TICHY, R. F.: *Sequences, Discrepancies and Applications*, Springer--Verlag, Berlin, 1997.

[5] HALTON, J. H.: *On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals*, Numer. Math. **2** (1960), 84–90.
[6] HELLEKALEK, P.: *A notion of diaphony based on p-adic arithmetics*, Acta Arith. **139** (2010), 117–129.
[7] HELLEKALEK, P.—KRITZER, P.—PILLICHSHAMMER, F.: *Open type quasi-Monte Carlo integration based on Halton sequences in weighted Sobolev spaces*, (2014), (submitted).
[8] KOROBOV, N. M.: *Number-theoretic methods in approximate analysis*, Gosudarstv. Izdat. Fiz.-Mat. Lit., Moscow, 1963. (In Russian)
[9] KRITZER, P.—PILLICHSHAMMER, F.: *On the component by component construction of polynomial lattice point sets for numerical integration in weighted Sobolev spaces.* Unif. Distrib. Theory **6** (2011), 79–100.
[10] KUIPERS, L.—NIEDERREITER, H.: *Uniform Distribution of Sequences*, John Wiley, New York, 1974.
[11] KUO, F. Y.: *Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces*, J. Complexity **19** (2003), 301–320.
[12] NIEDERREITER, H.: *Random Number Generation and Quasi-Monte Carlo Methods.* In: CBMS-NSF Series in Applied Mathematics Vol. 63. SIAM, Philadelphia, 1992.
[13] NOVAK, E.—WOŹNIAKOWSKI, H.: *Tractability of Multivariate Problems. Volume I: Linear Information.* In: EMS Tracts in Mathematics, Vol. 6. European Mathematical Society (EMS), Zürich, 2008.
[14] PILLICHSHAMMER, F.: *The p-adic diaphony of the Halton sequence.* Funct. Approx. Comment. Math. **49** (2013), 91–102.
[15] SLOAN, I. H.—JOE, S.: *Lattice Methods for Multiple Integration.* Oxford University Press, New York and Oxford, 1994.
[16] SLOAN, I. H.—REZTSOV, A. V.: *Component-by-component construction of good lattice rules*, Math. Comp. **71** (2002), 263–273.
[17] SLOAN, I. H.—WOŹNIAKOWSKI, H.: *When are quasi-Monte Carlo algorithms efficient for high-dimensional integrals?* J. Complexity **14**, (1998), 1–33.
[18] WANG, X.: *A constructive approach to strong tractability using quasi-Monte Carlo algorithms*, J. Complexity **18** (2002), 683–701.

Received January 29, 2015
Accepted April 16, 2015

Peter Kritzer
Friedrich Pillichshammer
Institut für Finanzmathematik und angewandte Zahlentheorie
Johannes Kepler Universität Linz
Altenbergerstraße 69
A-4040 Linz
AUSTRIA
E-mail: peter.kritzer(at)jku.at
friedrich.pillichshammer(at)jku.at

Uniform Distribution Theory **10** (2015), no.2, 67–75

uniform distribution theory

ON THE DISTRIBUTION OF POWERS OF REAL NUMBERS MODULO 1

Simon Baker

Dedicated to Professor Harald Niederreiter on the occasion of his 70th birthday

ABSTRACT. Given a strictly increasing sequence of positive real numbers tending to infinity $(q_n)_{n=1}^{\infty}$, and an arbitrary sequence of real numbers $(r_n)_{n=1}^{\infty}$. We study the set of $\alpha \in (1,\infty)$ for which $\lim_{n\to\infty} \|\alpha^{q_n} - r_n\| = 0$. In [3] Dubickas showed that whenever $\lim_{n\to\infty}(q_{n+1} - q_n) = \infty$, there always exists a transcendental α for which $\lim_{n\to\infty} \|\alpha^{q_n} - r_n\| = 0$. Adapting the approach of Bugeaud and Moshchevitin [2], we improve upon this result and show that whenever $\lim_{n\to\infty}(q_{n+1} - q_n) = \infty$, then for any interval $I \subset (1,\infty)$ the set of $\alpha \in I$ satisfying $\lim_{n\to\infty} \|\alpha^{q_n} - r_n\| = 0$ is of Hausdorff dimension 1.

Communicated by Arturas Dubickas

1. Introduction

It is a well known result of Koksma that for almost every $\alpha \in (1,\infty)$ the sequence $(\{\alpha^n\})_{n=1}^{\infty}$ is uniformly distributed modulo 1 [7]. Here and throughout almost every is meant with respect to the Lebesgue measure, and $\{\cdot\}$ denotes the fractional part of a real number. It is a long standing problem to determine the set of $\alpha \in (1,\infty)$ for which $\lim_{n\to\infty} \|\alpha^n\| = 0$, where $\|\cdot\|$ denotes the distance to the nearest integer. The only known examples of numbers satisfying this property are Pisot numbers, that is algebraic integers whose Galois conjugates all have modulus strictly less than 1. It was shown independently by Hardy [5] and Pisot [9], that if α is an algebraic number and $\lim_{n\to\infty} \|\alpha^n\| = 0$, then α is a Pisot number. Moreover, Pisot in [8] showed that there are only countably many $\alpha \in (1,\infty)$ satisfying $\lim_{n\to\infty} \|\alpha^n\| = 0$. This naturally leads to the question:

Is there a transcendental $\alpha \in (1,\infty)$ satisfying $\lim_{n\to\infty} \|\alpha^n\| = 0$?

2010 Mathematics Subject Classification: 11K06, 11K31.
Keywords: Powers of a real number, Uniform distribution.

This question is highly non trivial and currently out of our reach.

In this paper, instead of studying the distribution of the sequence $\{\alpha\}, \{\alpha^2\}, \{\alpha^3\}, \ldots$, we consider the distribution of the sequence $\{\alpha\}, \{\alpha^4\}, \{\alpha^9\}, \ldots$, or more generally $\{\alpha^{q_1}\}, \{\alpha^{q_2}\}, \{\alpha^{q_3}\}, \ldots$ where $(q_n)_{n=1}^{\infty}$ is a strictly increasing sequence of positive real numbers tending to infinity. We emphasise that the q_n are not necessarily natural numbers. We remark that if $\liminf_{n\to\infty}(q_{n+1}-q_n)>0$, then for almost every $\alpha \in (1,\infty)$ the sequence $(\{\alpha^{q_n}\})_{n=1}^{\infty}$ is uniformly distributed modulo 1. The proof of this statement is a simple adaptation of the proof of Theorem 1.10 in [1]. All of the sequences we will be considering shall satisfy $\liminf_{n\to\infty}(q_{n+1} - q_n) > 0$.

We are interested in the set of $\alpha \in (1,\infty)$ for which $(\{\alpha^{q_n}\})_{n=1}^{\infty}$ is not uniformly distributed modulo 1. In particular, the set of $\alpha \in (1,\infty)$ for which $\lim_{n\to\infty} \|\alpha^{q_n}\| = 0$, or more generally the set of $\alpha \in (1,\infty)$ which satisfy $\lim_{n\to\infty} \|\alpha^{q_n} - r_n\| = 0$, where $(r_n)_{n=1}^{\infty}$ is an arbitrary sequence of real numbers. Let

$$E(q_n, r_n) := \Big\{\alpha \in (1,\infty) : \lim_{n\to\infty} \|\alpha^{q_n} - r_n\| = 0\Big\}.$$

In [3], Dubickas showed that whenever $\lim_{n\to\infty}(q_{n+1} - q_n) = \infty$, then it is possible to construct a transcendental α contained in $E(q_n, r_n)$. Note that the α he constructs is always larger than 2. Our main result is the following.

Theorem 1.1. *Let $(q_n)_{n=1}^{\infty}$ be a strictly increasing sequence of positive real numbers satisfying $\lim_{n\to\infty}(q_{n+1} - q_n) = \infty$, and let $(r_n)_{n=1}^{\infty}$ be an arbitrary sequence of real numbers. Then for any interval $I \subset (1,\infty)$ the set of $\alpha \in I$ satisfying $\lim_{n\to\infty} \|\alpha^{q_n} - r_n\| = 0$ is of Hausdorff dimension 1.*

The proof we give of Theorem 1.1 is based upon the approach of Bugeaud and Moshchevitin [2], which in turn is based upon the approach of Vijayaraghavan [10]. They show that for any $\epsilon > 0$ and $(r_n)_{n=1}^{\infty}$ a sequence of real numbers, there exists a set of Hausdorff dimension 1 for which $\|\alpha^n - r_n\| < \epsilon$ for all $n \geq 1$. The set of $\alpha \in (1,\infty)$ which satisfy $\|\alpha^n - r_n\| < \epsilon$ for all n sufficiently large is studied further in [6].

Given $(\{\alpha^{q_n}\})_{n=1}^{\infty}$ is uniformly distributed modulo 1 for almost every $\alpha \in (1,\infty)$, Theorem 1.1 is somewhat surprising in that it states that there exists a set, which in some sense is as large as we could hope for, which exhibits completely the opposite behaviour of uniform distribution. Indeed, taking $(r_n)_{n=1}^{\infty}$ to be the constant sequence $r_n = \kappa$ for some $\kappa \in (0,1)$, then for any interval $I \subset (1,\infty)$ the set of $\alpha \in I$ satisfying $\lim_{n\to\infty}\{\alpha^{q_n}\} = \kappa$ is of Hausdorff dimension 1.

2. Proof of Theorem 1.1

We prove Theorem 1.1 via a Cantor set construction. To help our exposition we briefly recall some of the theory from [4] on this type of construction. Let $E_1 \subset \mathbb{R}$ be an arbitrary closed interval and $E_1 \supset E_2 \supset E_3 \supset \cdots$ be a decreasing sequence of sets, where each E_n is a finite union of disjoint closed intervals, where each element of E_n contains at least two elements of E_{n+1}, and the maximum length of the intervals in E_n tends to 0 as $n \to \infty$. Then the set

$$E = \bigcap_{n=1}^{\infty} E_n \tag{1}$$

is the Cantor set associated to the sequence $(E_n)_{n=1}^{\infty}$. The following proposition appears at Example 4.6 in [4].

Proposition 2.1. *Suppose in the construction of E above each interval in E_{n-1} contains at least m_n intervals of E_n which are separated by gaps of at least γ_n, where $0 < \gamma_{n+1} < \gamma_n$ for each n. Then*

$$\dim_H(E) \geq \liminf_{n\to\infty} \frac{\log m_1 \cdots m_{n-1}}{-\log m_n \gamma_n}.$$

We are now is a position to prove Theorem 1.1.

Proof of Theorem 1.1. We begin by fixing $\lambda \in (1, \infty)$, $\delta > 0$ some small positive constant, and let $(r_n)_{n=1}^{\infty}$ be our sequence of real numbers. Without loss of generality we may assume that $r_n \in [-1/2, 1/2)$ for all $n \in \mathbb{N}$. To prove our result it is sufficient to prove that $[\lambda, \lambda+\delta] \cap E(q_n, r_n)$ is of Hausdorff dimension 1.

(1) Replacing $(q_n)_{n=1}^{\infty}$ with $(\tilde{q}_n)_{n=1}^{\infty}$. Let $\epsilon > 0$ be some small positive constant. We now replace our sequence $(q_n)_{n=1}^{\infty}$ with $(\tilde{q}_n)_{n=1}^{\infty}$, and our sequence $(r_n)_{n=1}^{\infty}$ with $(\tilde{r}_n)_{n=1}^{\infty}$. We will pick our $(\tilde{q}_n)_{n=1}^{\infty}$ and $(\tilde{r}_n)_{n=1}^{\infty}$ in such a way that $E(\tilde{q}_n, \tilde{r}_n) \subset E(q_n, r_n)$. We then use Proposition 2.1 to determine a lower bound for $\dim_H(E(\tilde{q}_n, \tilde{r}_n) \cap [\lambda, \lambda + \delta])$, which in turn provides a lower bound for $\dim_H(E(q_n, r_n) \cap [\lambda, \lambda + \delta])$. The feature of the sequence $(\tilde{q}_n)_{n=1}^{\infty}$ that we will exploit in our proof, is that this new sequence does not grow too fast, yet importantly we still have $\lim_{n\to\infty}(\tilde{q}_{n+1} - \tilde{q}_n) = \infty$. The sequence $(\tilde{q}_n)_{n=1}^{\infty}$ and the rate at which we control the growth of $(\tilde{q}_n)_{n=1}^{\infty}$ shall depend on ϵ. For ease of exposition we drop the dependence of $(\tilde{q}_n)_{n=1}^{\infty}$ on ϵ from our notation.

We begin our construction by asking whether

$$q_{n+1} \leq (1 + \epsilon) q_n \tag{2}$$

is satisfied for all $n \in \mathbb{N}$. If it is, we set $(q_n)_{n=1}^{\infty} = (\tilde{q}_n)_{n=1}^{\infty}$, $(r_n)_{n=1}^{\infty} = (\tilde{r}_n)_{n=1}^{\infty}$ and stop. Suppose our sequence $(q_n)_{n=1}^{\infty}$ does not satisfy (2) for all $n \in \mathbb{N}$. Let $N \in \mathbb{N}$ be the first $n \in \mathbb{N}$ for which (2) fails. We now introduce additional terms in our sequence $(q_n)_{n=1}^{\infty}$, situated between q_N and q_{N+1} at

$$\tilde{q}_N^j := q_N + j\Big(\frac{\epsilon q_N}{2}\Big) \quad \text{for } \ j = 1, \dots, m.$$

Here m is the smallest natural number for which $q_N + m(\frac{\epsilon q_N}{2}) \in [q_{N+1} - \epsilon q_N, q_{N+1}]$. To each $\tilde{q}_N^j$ we associate an arbitrary real number $\tilde{r}_N^j \in [-1/2, 1/2)$, these terms are then placed within the sequence $(r_n)_{n=1}^{\infty}$ between r_N and r_{N+1}. Importantly the elements q_N and q_{N+1} are still placed in the positions corresponding to r_N and r_{N+1}.

The following inequalities are straightforward consequences of our construction

$$\begin{aligned} \tilde{q}_N^1 &\leq (1+\epsilon) q_n \\ \tilde{q}_N^{j+1} &\leq (1+\epsilon)\tilde{q}_N^j \quad \text{for } \ j = 1, \dots, m-1, \\ q_{N+1} &\leq (1+\epsilon)\tilde{q}_N^m. \end{aligned} \tag{3}$$

In other words, all of the new terms in our sequences satisfy (2). The new terms in our sequence also satisfy

$$\begin{aligned} \tilde{q}_N^1 - q_N &= \frac{\epsilon q_N}{2} \\ \tilde{q}_N^{j+1} - \tilde{q}_N^j &= \frac{\epsilon q_N}{2} \quad \text{for } \ j = 1, \dots, m-1, \\ q_{N+1} - \tilde{q}_N^m &\geq \frac{\epsilon q_N}{2}. \end{aligned} \tag{4}$$

So if N was large the gaps between successive terms in our sequences would be large. This property is what allows us to ensure $\lim_{n\to\infty}(\tilde{q}_{n+1} - \tilde{q}_n) = \infty$.

We now take our new sequence and ask if it satisfies (2) for all $n \in \mathbb{N}$. If it does, then our construction is complete, and we set $(\tilde{q}_n)_{n=1}^{\infty}$ and $(\tilde{r}_n)_{n=1}^{\infty}$ to be our new sequences. If not, we find the smallest n for which it fails and repeat the above steps. Repeating this process indefinitely if necessary, we construct sequences $(\tilde{q}_n)_{n=1}^{\infty}$ and $(\tilde{r}_n)_{n=1}^{\infty}$ for which

$$\tilde{q}_{n+1} \leq (1+\epsilon)\tilde{q}_n, \tag{5}$$

holds for all $n \in \mathbb{N}$, we retain the property

$$\lim_{n\to\infty} (\tilde{q}_{n+1} - \tilde{q}_n) = \infty, \tag{6}$$

and

$$E(\tilde{q}_n, \tilde{r}_n) \subset E(q_n, r_n). \tag{7}$$

The fact that (6) holds is a consequence of (4). The final property (7) holds because the original terms in our sequence $(q_n)_{n=1}^{\infty}$ keep their corresponding r_n, and at each step in our construction we only ever introduced finitely many terms between a q_n and a q_{n+1}. So if α satisfies $\|\alpha^{\tilde{q}_n} - \tilde{r}_n\| \to 0$ then it also satisfies $\|\alpha^{q_n} - r_n\| \to 0$, i. e., (7) holds.

(2) Construction of our Cantor set. We now construct our Cantor set E. Our set E will be contained in $[\lambda, \lambda+\delta] \cap E(\tilde{q}_n, \tilde{r}_n)$ and we will be able to use Proposition 2.1 to obtain estimates on $\dim_H(E)$. We let

$$\epsilon_n := \frac{1}{2(\tilde{q}_{n+1} - \tilde{q}_n)},$$

by (6) we have $\epsilon_n \to 0$. Let us fix $\eta \in (0,1)$ some parameter that we will eventually let tend to 1. Let N be sufficiently large that

$$2\epsilon_n \lambda^{\tilde{q}_{n+1}-\tilde{q}_n} = \frac{\lambda^{\tilde{q}_{n+1}-\tilde{q}_n}}{\tilde{q}_{n+1}-\tilde{q}_n} \geq \lceil \lambda^{\eta(\tilde{q}_{n+1}-\tilde{q}_n)} \rceil + 2 \quad \text{for all } \ n \geq N. \tag{8}$$

We may also assume that this N is sufficiently large that

$$(\lambda+\delta)^{\tilde{q}_N} - \lambda^{\tilde{q}_N} \geq 4 \quad \text{and} \quad \epsilon_n < 1/2 \quad \text{for all } \ n \geq N.$$

We let

$$a_n := \tilde{r}_n - \epsilon_n \quad \text{and} \quad b_n := \tilde{r}_n + \epsilon_n \quad \text{for all } \ n \in \mathbb{N}.$$

By our assumptions on $\tilde{r}_n$ and ϵ_n, we may assume that $a_n, b_n \in (-1,1)$ for all $n \geq N$.

Since $(\lambda+\delta)^{\tilde{q}_N} - \lambda^{\tilde{q}_N} \geq 4$, there exists an integer j_N for which $j_N, j_N+1, \ldots, j_N+m+1$ are contained in $[\lambda^{\tilde{q}_N}, (\lambda+\delta)^{\tilde{q}_N}]$, where m is some natural number greater than or equal to 2. We ignore the first and the last of these integers and focus on $j_N+1, \ldots j_N+m$. To each of these integers $h = j_N+1, \ldots j_N+m$ we associate the interval

$$I_{N,h} := [(h+a_N)^{1/\tilde{q}_N}, (h+b_N)^{1/\tilde{q}_N}].$$

By our construction each $I_{N,h}$ is contained in $[\lambda, \lambda+\delta]$, and each $\alpha \in I_{N,h}$ satisfies $\|\alpha^{\tilde{q}_N} - \tilde{r}_N\| < \epsilon_N$. Let E_N be the set of all intervals $I_{N,h}$. For each h we have

$$\begin{aligned}(h+b_N)^{\tilde{q}_{N+1}/\tilde{q}_N} - (h+a_N)^{\tilde{q}_{N+1}/\tilde{q}_N} &\geq \Big((h+b_N)-(h+a_N)\Big)(h+a_N)^{\frac{\tilde{q}_{N+1}-\tilde{q}_N}{\tilde{q}_N}} \\ &\geq 2\epsilon_N \lambda^{\tilde{q}_{N+1}-\tilde{q}_N} \\ &\geq \lceil \lambda^{\eta(\tilde{q}_{N+1}-\tilde{q}_N)} \rceil + 2.\end{aligned} \tag{9}$$

Where the last inequality is by (8). Therefore there exists an integer j_{N+1} such that $j_{N+1}, j_{N+1}+1, \dots, j_{N+1}+\lceil \lambda^{\eta(\tilde{q}_{n+1}-\tilde{q}_n)} \rceil + 1$ are all contained in

$$[(h+a_N)^{\tilde{q}_{n+1}/\tilde{q}_n}, (h+b_N)^{\tilde{q}_{n+1}/\tilde{q}_n}].$$

To each $h = j_{N+1}+1, \dots, j_{N+1}+\lceil \lambda^{\eta(\tilde{q}_{n+1}-\tilde{q}_n)} \rceil$ we associate the interval

$$I_{N+1,h} = [(h+a_{N+1})^{1/\tilde{q}_{N+1}}, (h+b_{N+1})^{1/\tilde{q}_{N+1}}].$$

Importantly each interval $I_{N+1,h}$ is contained in an element of E_N, and this element contains precisely $m_N := \lceil \lambda^{\eta(\tilde{q}_{N+1}-\tilde{q}_N)} \rceil$ of these intervals. We let E_{N+1} denote the set of $I_{N+1,h}$. Any $\alpha \in I_{N+1,H}$ is contained in $[\lambda, \lambda+\delta]$ and satisfies

$$\|\alpha^{\tilde{q}_N} - \tilde{r}_N\| < \epsilon_N \quad \text{and} \quad \|\alpha^{\tilde{q}_{N+1}} - \tilde{r}_{N+1}\| < \epsilon_{N+1}.$$

We may show that (9) holds with $a_N, b_N, \tilde{q}_N, \tilde{q}_{N+1}$ replaced by a_{N+1}, b_{N+1}, $\tilde{q}_{N+1}$, $\tilde{q}_{N+2}$, and we may therefore repeat the above steps accordingly. Moreover, we may repeat the procedure described above for every subsequent n. To each $n \geq N$ we let E_n denote the set of interval $I_{n,h}$ produced in our construction. The following properties follow from our construction:

- Let $I_{n,h} \in E_n$, then for each $\alpha \in I_{n,h}$ we have
$$\|\alpha^{\tilde{q}_i} - \tilde{r}_i\| < \epsilon_n \quad \text{for } \ i = N, N+1, \dots, n.$$
- $E_n \subset E_{n-1} \subset \cdots \subset E_N$.
- $E_n \subset [\lambda, \lambda+\delta]$.

If we let

$$E = \bigcap_{n=N}^{\infty} E_n,$$

it is clear that any $x \in E$ is contained in $[\lambda, \lambda+\delta]$, and satisfies $\|x^{\tilde{q}_n} - \tilde{r}_n\| < \epsilon_n$ for all $n \geq N$, so

$$E \subset [\lambda, \lambda+\delta] \cap E(\tilde{q}_n, \tilde{r}_n).$$

It is a consequence of our construction that each element of E_n contains exactly

$$m_n := \lceil \lambda^{\eta(\tilde{q}_{n+1}-\tilde{q}_n)} \rceil \tag{10}$$

elements of E_{n+1}. It may also be shown that the distance between any two intervals in E_n is always at least

$$\gamma_n := \frac{c}{\tilde{q}_n(\lambda+\delta)^{\tilde{q}_n-1}}, \tag{11}$$

where c is some positive constant that is independent of n.

Applying Proposition 2.1, combined with (10) and (11), we obtain the following bounds on the Hausdorff dimension of E:

$$\begin{aligned}
\dim_H(E) &\geq \liminf_{n\to\infty} \frac{\log m \cdot m_N \cdots m_{n-1}}{-\log m_n \gamma_n} \\
&\geq \liminf_{n\to\infty} \frac{\log 2 \cdot \lceil \lambda^{\eta(\tilde{q}_{N+1}-\tilde{q}_N)} \rceil \cdots \lceil \lambda^{\eta(\tilde{q}_n-\tilde{q}_{n-1})} \rceil}{-\log \frac{c\lceil \lambda^{\eta(\tilde{q}_{n+1}-\tilde{q}_n)} \rceil}{\tilde{q}_n(\lambda+\delta)^{\tilde{q}_n-1}}} \\
&\geq \liminf_{n\to\infty} \frac{\eta(\tilde{q}_n-\tilde{q}_N)\log\lambda + \log 2}{-\log \frac{c\lceil \lambda^{\eta(\tilde{q}_{n+1}-\tilde{q}_n)} \rceil}{\tilde{q}_n(\lambda+\delta)^{\tilde{q}_n-1}}} \\
&\geq \liminf_{n\to\infty} \frac{\eta(\tilde{q}_n-\tilde{q}_N)\log\lambda + \log 2}{-\log\lceil \lambda^{\eta(\tilde{q}_{n+1}-\tilde{q}_n)} \rceil - \log c + (\tilde{q}_n-1)\log(\lambda+\delta) + \log \tilde{q}_n} \\
&\geq \liminf_{n\to\infty} \frac{\eta(\tilde{q}_n-\tilde{q}_N)\log\lambda + \log 2}{-\eta(\tilde{q}_{n+1}-\tilde{q}_n)\log\lambda - \log c + (\tilde{q}_n-1)\log(\lambda+\delta) + \log \tilde{q}_n} \\
&\geq \frac{\eta\log\lambda}{\eta\epsilon\log\lambda + \log(\lambda+\delta)}.
\end{aligned}$$

Since η was arbitrary we may let it converge to 1 so

$$\dim_H\big([\lambda,\lambda+\delta]\cap E(\tilde{q}_n,\tilde{r}_n)\big) \geq \frac{\log\lambda}{\epsilon\log\lambda+\log(\lambda+\delta)}.$$

Therefore, by (7)

$$\dim_H\big([\lambda,\lambda+\delta]\cap E(q_n,r_n)\big) \geq \frac{\log\lambda}{\epsilon\log\lambda+\log(\lambda+\delta)},$$

but since ϵ is arbitrary we may conclude that

$$\dim_H\big([\lambda,\lambda+\delta]\cap E(q_n,r_n)\big) \geq \frac{\log\lambda}{\log(\lambda+\delta)}.$$

The argument we have presented also works for any $\delta' \in (0,\delta)$ this implies

$$\dim_H\big([\lambda,\lambda+\delta']\cap E(q_n,r_n)\big) \geq \log\lambda/\log(\lambda+\delta').$$

Moreover

$$[\lambda,\lambda+\delta']\cap E(q_n,r_n) \ \subset [\lambda,\lambda+\delta]\cap E(q_n,r_n),$$

so

$$\dim_H\big([\lambda,\lambda+\delta]\cap E(q_n,r_n)\big) \geq \dim_H\big([\lambda,\lambda+\delta']\cap E(q_n,r_n)\big) \geq \frac{\log\lambda}{\log(\lambda+\delta')}.$$

Letting δ' tend to zero we deduce that $\dim_H\big([\lambda,\lambda+\delta]\cap E(q_n,r_n)\big) = 1$. □

We conclude with a few remarks on our proof and the speed at which $\|\alpha^{q_n} - r_n\|$ converges to zero. In our proof of Theorem 1.1 we set

$$\epsilon_n = 1/2(\tilde{q}_{n+1} - \tilde{q}_n).$$

This choice of ϵ_n is somewhat arbitrary, our proof still works with any sequence ϵ_n which tends to zero, as long as for any $\eta \in (0,1)$ we have

$$2\epsilon_n \lambda^{\tilde{q}_{n+1}-\tilde{q}_n} \geq \lceil \lambda^{\eta(\tilde{q}_{n+1}-\tilde{q}_n)} \rceil + 2$$

for all n sufficiently large.

If $(q_n)_{n=1}^{\infty}$ satisfies $\lim_{n\to\infty} q_{n+1}/q_n = 1$, then it is not necessary to introduce the sequences $(\tilde{q}_n)_{n=1}^{\infty}$ and $(\tilde{r}_n)_{n=1}^{\infty}$ in the proof of Theorem 1.1. This means we can say something about the speed of convergence. If ϵ_n decays to zero sufficiently slowly that for any $\eta \in (0,1)$ and $\lambda \in (1,\infty)$, we have

$$2\epsilon_n \lambda^{q_{n+1}-q_n} \geq \lceil \lambda^{\eta(q_{n+1}-q_n)} \rceil + 2,$$

for all n sufficiently large. Then the argument given in the proof of Theorem 1.1 yields a set of Hausdorff dimension 1 within any interval satisfying

$$\|\alpha^{q_n} - r_n\| = O(\epsilon_n).$$

As an example, for any $k \in \mathbb{N}$ there exists a set of Hausdorff dimension 1 within any interval satisfying

$$\|\alpha^{n^2}\| = O(n^{-k}).$$

Acknowledgements. The author is grateful to Yann Bugeaud for pointing out [3] and [6], and for some initial feedback.

REFERENCES

[1] BUGEAUD, Y.: *Distribution modulo one and Diophantine approximation,* Cambridge Tracts in Mathematics, 193. Cambridge University Press, Cambridge, 2012.

[2] BUGEAUD, Y.—MOSHCHEVITIN, V.: *On fractional parts of powers of real numbers close to* 1, Math. Z. 271 (2012), no. 3–4, 627–637.

[3] DUBICKAS, A.: *On the powers of some transcendental numbers,* Bull. Austral. Math. Soc. **76** (2007), no. 3, 433–440.

[4] FALCONER, K.: *Mathematical Foundations and Applications,* John Wiley & Sons, Ltd., Chichester, 2014.

[5] HARDY, G. H.: *A problem of Diophantine approximation,* J. Indian Math. Soc. **11** (1919), 162–166.

[6] KAHANE, J.-P.: *Sur la répartition des puissances modulo 1,* C. R. Math. Acad. Sci. Paris **352** (2014), no. 5, 383–385.

[7] KOKSMA, J. F.: *Ein mengentheoretischer Satz über die Gleichverteilung modulo Eins,* Compositio Math. **2** (1935), 250–258.

[8] PISOT, C.: *Sur la répartition modulo* 1 *des puissances successives d'un même nombre,* C.R. Acad. Sci. Paris **204** (1937), 312–314.
[9] PISOT, C.: *La répartition modulo 1 et les nombres algébriques,* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (2) **7** (1938), no. 3–4, 205–248.
[10] VIJAYARAGHAVAN, T.: *On the fractional parts of powers of a number. IV,* J. Indian Math. Soc. (N.S.) **12**, (1948). 33–39.

Received November 24, 2014
Accepted May 7, 2015

Simon Baker
School of Mathematics
The University of Manchester
Oxford Road
Manchester
M13 9PL
UNITED KINGDOM
E-mail: simonbaker412@gmail.com

Uniform Distribution Theory **10** (2015), no.2, 77–91

uniform distribution theory

STABILITY OF BALANCING SEQUENCE MODULO p

Sudhansu Sekhar Rout—Ravi Kumar Davala—Gopal Krishna Panda

Dedicated to Professor Harald Niederreiter on the occasion of his 70th birthday

ABSTRACT. Stability is an important aspect of a number sequence. It is known that Fibonacci sequence is stable modulo 2 and 5. The objective of the paper is to study the stability of the balancing sequence modulo primes.

Communicated by András Sárközy

1. Introduction

As introduced by Behera and Panda [1], balancing numbers x and balancers r are solutions of the Diophantine equation

$$1 + 2 + 3 + \cdots + (x-1) = (x+1) + (x+2) + \cdots + (x+r). \qquad (1.1)$$

As a consequence of (1.1), if x is a balancing number then $x^2 = \frac{(x+r)(x+r+1)}{2}$ is a triangular number or equivalently, $8x^2+1$ is a perfect square and $\sqrt{8x^2+1}$ is called a Lucas-balancing number [8]. Writing $8x^2 + 1 = y^2$, we are lead to the Pell's equation $y^2 - 8x^2 = 1$ satisfied by the Lucas-balancing and balancing numbers. The n^{th} balancing and Lucas-balancing number are denoted by B_n and C_n respectively. The balancing numbers satisfy the recurrence relation $B_0 = 0, B_1 = 1$ and $B_{n+1} = 6B_n - B_{n-1}$ for $n \geq 2$. The sequence of balancing numbers modulo m is periodic and the period modulo m is denoted by $\pi(m)$ [9]. By definition, $\pi(m)$ is the smallest natural number to satisfy $B_{\pi(m)} \equiv 0, B_{\pi(m)+1} \equiv 1 \pmod m)$. The computation of $\pi(m)$ depends on the factorization of m, but for arbitrary primes p, there is no exact formula for $\pi(p)$, though certain divisibility relations for $\pi(p)$ are known [9]. The rank of apparition or simply rank of balancing sequence modulo m is the least positive

2010 Mathematics Subject Classification: 11A06, 11A07.
Keywords: balancing number, uniform distribution modulo m, stability, prime.

integer r such that $B_r \equiv 0 \pmod{m}$ and let it be denoted by $\alpha(m)$. Thus, $\alpha(m)$ is the index of the first non-zero balancing number B_n which is divisible by m. Niven [7] introduced the notion of uniform distribution of a sequence of integer as follows. A sequence of integers $\mathcal{A} = \{a_n : n = 0, 1, \ldots\}$ is called uniformly distributed modulo $m \geq 2$ if

$$\lim_{N\to\infty} \frac{1}{N} \#\{n < N : a_n \equiv b \pmod{m}\} = \frac{1}{m}$$

for any $b \in \{0, 1, \ldots, m-1\}$.

For a fixed modulus m and feasible residue r, we denote the number of occurrences of r in a period of the balancing sequence by $\nu_B(m, r)$. This function is the frequency distribution function of balancing sequence modulo m. In the early 1970's, interest in the distribution functions of binary recurrence sequences centered on the characterization of those sequences that have constant frequency distribution function, i.e., sequences that are uniformly distributed. Denoting by

$$\Omega_B(m) := \{\nu_B(m, r) : r \in \{0, 1, \ldots, m-1\}\} \setminus \{0\}, \tag{1.2}$$

the set of all frequencies of the feasible residues modulo m in a period, the balancing sequence is uniformly distributed whenever $\#\{\Omega_B(m)\} = 1$. Stability of the balancing sequence comes into picture when $\#\{\Omega_B(m)\}$ is not constant and this generalize the concept of uniform distribution and also the notion of f-uniform distribution modulo prime powers [11]. The concrete and precise definition of stability was due to Carlip and Jacobson [3]. We prefer to state this definition for the balancing sequence, though it can be stated for any arbitrary sequence.

Definition 1.1. The balancing sequence is said to be stable modulo a prime p if there is a positive integer N such that $\Omega_B(p^k) = \Omega_B(p^N)$ for all $k \geq N$.

For better understanding of the above concept, the following examples will be helpful.

Example 1.2. Consider the balancing sequence modulo $3, 9, 27$. It is easy to see that

$$\nu_B(3,0) = 2,\ \nu_B(3,1) = 1 = \nu_B(3,2),$$
$$\nu_B(9,0) = \nu_B(9,3) = \nu_B(9,6) = 2,\ \nu_B(9,1) = \nu_B(9,8) = 3,$$
$$\nu_B(27,1) = \nu_B(27,26) = 6, \nu_B(27,8) = \nu_B(27,19) = 3,$$
$$\nu_B(27,i) = 2 \text{ for } i \equiv 0, \pm 3, \pm 6, \pm 9, \pm 12 \pmod{27}.$$

Therefore

$$\Omega_B(3) = \{1, 2\},\ \ \Omega_B(3^2) = \{2, 3\},\ \ \Omega_B(3^3) = \{2, 3, 6\}. \tag{1.3}$$

From (1.3), we observe that the elements in the set $\Omega_B(3^k)$ increases as k increases.

Example 1.3.

$$\nu_B(5,0) = 2 = \nu_B(5,1) = \nu_B(5,4),$$

$$\nu_B(25,i) = 2 \text{ for } i \equiv 0, \pm 1, \pm 4, \pm 5, \pm 6, \pm 10, \pm 14, \pm 16 \pmod{25}.$$

Hence

$$\Omega_B(5) = \{2\} = \Omega_B(5^2). \tag{1.4}$$

Equation (1.3) is an indication that the balancing sequence may not be stable modulo 3, while (1.4) shows the possible stability of the sequence modulo 5.

Bundschuh [2] studied the stability of Lucas sequence modulo 2 and 5 and found that the sequence is not stable for these two primes. Somer and Carlip [10] demonstrated several classes of binary recurrences which are not p-stable and established sufficient criteria for such recurrences to be p-stable. We will show that the balancing sequence is stable for two particular classes of primes. In this paper, we will completely describe the function $\nu_B(p^k, \cdot)$. We will show that $\nu_B(2^k, \cdot) = \{1\}$ and hence $\Omega_B(2^k) = \{1\}, \nu_B(p^k, \cdot) = \{1\}$ when $p \equiv -1 \pmod 8$ and $\nu_B(p^k, \cdot) = \{2\}$ when $p \equiv -3 \pmod 8$. These results would confirm the stability of the balancing sequence modulo p when $p \equiv -1$, $-3 \pmod 8$. Finally we have shown that balancing sequence is not stable modulo primes $p \equiv 3 \pmod 8$. However, for some primes $p \equiv 1 \pmod 8$ the balancing sequence is stable.

2. Preliminaries

In this section, we present some results which will be needed in the sequel. Throughout the remaining part of this paper, p represents an odd prime. For any non-zero integer a, $\text{ord}_p a = m$ if $p^m \mid a$ but $p^{m+1} \nmid a$. Important properties of ord_p are $\text{ord}_p(ab) = \text{ord}_p(a) + \text{ord}_p(b), \text{ord}_p(a+b) \geq \min(\text{ord}_p(a), \text{ord}_p(b))$ [5]. Thus $a \equiv b \pmod{p^k}$ is equivalent to $\text{ord}_p(a-b) \geq k$.

We also need the following results relating to periods of balancing numbers.

Theorem 2.1. ([9]) *For any natural number $n > 1$, $\pi(n) = n$ if and only if $n = 2^k$ for any $k \in \mathbb{N}$.*

Theorem 2.2. ([9]) *For any odd prime p, $\pi(p)$ divides $p-1$ if $p \equiv \pm 1 \pmod 8$ and $\pi(p)$ divides $p+1$ if $p \equiv \pm 3 \pmod 8$. Thus, if p is an odd prime, then $\pi(p)$ divides $p^2 - 1$.*

Lemma 2.3. *If the integers m and n are of the same parity, then*

$$B_m - B_n = 2B_{\frac{m-n}{2}}\, C_{\frac{m+n}{2}}, \tag{2.1}$$

$$C_m - C_n = 16B_{\frac{m-n}{2}} B_{\frac{m+n}{2}}. \tag{2.2}$$

Proof. It is well known that $B_{x\pm y} = B_x C_y \pm C_x B_y$ [8]. Thus $B_{x+y} - B_{x-y} = 2B_y C_x$; and taking $x + y = m, x - y = n$, we get $B_m - B_n = 2B_{\frac{m-n}{2}} C_{\frac{m+n}{2}}$. Similarly by virtue of the formula $C_{x\pm y} = C_x C_y \pm 8B_x B_y$, [8] $C_m - C_n = 16B_{\frac{m-n}{2}} B_{\frac{m+n}{2}}$ follows. □

Lemma 2.4. *If $B_n \equiv 0 \pmod p$, then $B_{2n} \equiv 0 \pmod p$ and $B_{2n+1} \equiv 1 \pmod p$.*

Proof. Since $B_n \equiv 0 \pmod p, B_{2n} = 2B_n C_n \equiv 0 \pmod p$ and $B_{2n+1} = B_n \cdot B_{n+2} - B_{n-1} \cdot B_{n+1} \equiv -(B_n^2 - 1) \equiv 1 \pmod p$ since $B_{n+1}B_{n-1} = B_n^2 - 1$, (see [8]). □

Lemma 2.5. *For any prime p, $\pi(p) = \alpha(p)$ or $\pi(p) = 2\alpha(p)$.*

Proof. Since $B_{\alpha(p)} \equiv 0 \pmod p$ by Lemma 2.4 we get $B_{2\alpha(p)} \equiv 0 \pmod p$ and $B_{2\alpha(p)+1} \equiv 1 \pmod p$. Thus, $\pi(p) | 2\alpha(p)$ and hence

$$\pi(p) = \alpha(p) \quad \text{or} \quad \pi(p) = 2\alpha(p). \qquad \square$$

Lemma 2.6. *If $\alpha(p^2) \neq \alpha(p)$ then $\alpha(p^l) = p^{l-1}\alpha(p)$. Further, if k is the largest integer such that $\alpha(p^k) = \alpha(p)$ and $l > k$, then $\alpha(p^l) = p^{l-k}\alpha(p)$*

Proof. The congruence $B_{\alpha(p^l)} \equiv 0 \pmod{p^l}$ gives $B_{\alpha(p^l)} = kp^l$ for some natural number k. By De-Moivre's Theorem for balancing numbers ([8])

$$C_{p\alpha(p^l)} + \sqrt{8}B_{p\alpha(p^l)} = \left(C_{\alpha(p^l)} + \sqrt{8}B_{\alpha(p^l)}\right)^p.$$

Hence, for $l > 1$

$$\begin{aligned} B_{p\alpha(p^l)} &= k\binom{p}{1} C_{\alpha(p^l)}^{p-1} p^l + 8k^3 \binom{p}{3} C_{\alpha(p^l)}^{p-3} p^{3l} + \cdots + 8^{\frac{p-1}{2}} k^p p^{pl} \\ &\equiv 0 \pmod{p^{l+1}}. \end{aligned} \tag{2.3}$$

It is clear from above equation that $\alpha(p^{l+1})$ divides $p\alpha(p^l)$. Since $\alpha(p^l)$ divides $\alpha(p^{l+1})$, it follows that $\alpha(p^{l+1}) = \alpha(p^l)$ or $\alpha(p^{l+1}) = p\alpha(p^l)$. For $l = 1$, the conclusion is that $\alpha(p^2) = \alpha(p)$ or $\alpha(p^2) = p\alpha(p)$; so if $\alpha(p^2) \neq \alpha(p)$, then $\alpha(p^2) = p\alpha(p)$. Continuing in this process we will arrive at $\alpha(p^l) = p^{l-1}\alpha(p)$. Further, if k is the largest integer such that $\alpha(p^k) = \alpha(p)$, then $\alpha(p^{k+t}) = p\alpha(p^{k+t-1}) = \cdots = p^t\alpha(p^k) = p^t\alpha(p)$ for each natural number t. □

The following lemma, which relates the order of B_n with order of n, will play a crucial role.

Lemma 2.7. *If $n \in \mathbb{N}$ and p is any arbitrary prime then $\alpha(p) \mid n$ if and only if $p \mid B_n$. Furthermore, if $\alpha(p) \mid n$, then*

$$\operatorname{ord}_p B_n \geq 1 + \operatorname{ord}_p n. \tag{2.4}$$

Proof. The proof of first part follows directly from the definition of $\alpha(p)$. Let $\operatorname{ord}_p B_n = t$ and $\operatorname{ord}_p n = s$. Then $n = kp^s$ where $\gcd(k, p) = 1$, and $\alpha(p) \mid n$ implies $\alpha(p) \mid kp^s$. Since $\alpha(p) \mid p^2 - 1$, by Theorem 2.2, $\gcd(\alpha(p), p) = 1$. Thus $\alpha(p)|k$ which gives $k = a\alpha(p)$ for some integer a. Putting the value of k in n, we get $n = a\alpha(p)p^s$. By definition, $p^t \| B_n$ if and only if $\alpha(p^t) \| n$. If $\alpha(p) \neq \alpha(p^2)$ then by Lemma 2.6, $p^{t-1}\alpha(p) \| n$. Putting the value of n, we get $p^{t-1}\alpha(p) \| a\alpha(p)p^s$ which implies $p^{t-1} \| a \cdot p^s$. Since $\gcd(k, p) = 1$ and $k = a\alpha(p)$, we have $\gcd(a, p) = 1$. Therefore $t - 1 = s$. If $m > 1$ is the largest integer such that $\alpha(p^m) = \alpha(p)$, then $p^{t-m}\alpha(p) \| n$ and proceeding as above we will reach at $t - m = s$. Hence combining both the cases we conclude that $t \geq 1 + s$. □

Similar results also hold for the Lucas balancing sequence. The proof of the following lemma is similar to that of Lemma 2.7 and is omitted.

Lemma 2.8. *Let $n \in \mathbb{N}$ and for any prime $p \equiv 3 \pmod 8$ define*

$$\beta(p) = \min\{r : C_r \equiv 0 \pmod p\}.$$

Then

$$\beta(p) \mid n \Leftrightarrow p \mid C_n \quad \text{and} \quad \beta(p) \mid n \Rightarrow \operatorname{ord}_p C_n = 1 + \operatorname{ord}_p n\,. \tag{2.5}$$

3. Stability of balancing sequence modulo 2

The Fibonacci sequence is stable modulo 2 and 5 [4]. In this section, we will show that the balancing sequence is also stable modulo 2.

Theorem 3.1. *$\nu_B(2^k, b) = 1$ for every residue b modulo 2^k and for any $k \in \mathbb{N}$.*

Proof. By virtue of Theorem 2.1, for $n > 1$, $\pi(n) = n$ if and only if $n = 2^k$ for any $k \in \mathbb{N}$. Using this result, we will show that each residue $b \in \{0, 1, \cdots, 2^k - 1\}$ occurs only once in a period modulo 2^k. Since B_n is even or odd according as n is even or odd, it follows that the least residue of $B_n, 0 \leq n \leq 2^k - 1$ modulo 2^k is also even or odd according as n is even or odd. To complete the proof, we have to show that no two least residue of $B_n, 0 \leq n \leq 2^k - 1$ are congruent modulo 2^k. Since B_{2m+1} and B_{2n} are incongruent modulo 2^k, it is sufficient to show that

$$B_{2m+1} \not\equiv B_{2n+1} \pmod{2^k} \text{ for } 0 < 2m + 1 < 2n + 1 < 2^k, \tag{3.1}$$

and

$$B_{2i} \not\equiv B_{2j} \pmod{2^k} \text{ for } 0 \leq 2i < 2j \leq 2^k. \tag{3.2}$$

Since $\pi(2^k) = 2^k$, it follows that $2^k \mid B_n$ if and only if $2^k \mid n$. Let us assume the contrary of (3.1), i. e.,

$$B_{2m+1} \equiv B_{2n+1} \pmod{2^k} \quad \text{for } 0 < 2m+1 < 2n+1 < 2^k,$$

hence 2^k divides $B_{2n+1} - B_{2m+1}$. Using (2.1), $2^k \mid 2B_{n-m}C_{m+n+1}$ implies $2^{k-1} \mid B_{n-m}$ as $\gcd(2, C_x) = 1$ for any natural number x. It easily follows from induction on k that if $2^{k-1} \mid B_{n-m}$, then $2^{k-1} \mid n-m$ which is a contradiction since $n - m < 2^{k-1}$. Thus (3.1) holds. In a similar fashion, (3.2) can be proved. □

From equation (1.2), we have $\Omega_B(2^k) = \{1\}$. The following corollary, which ascertains the stability of balancing sequence modulo 2, is a consequence of the above theorem.

Corollary 3.2. *The balancing sequence is stable modulo* 2.

4. Stability of balancing sequence modulo primes $p \equiv -1, -3 \pmod 8$

In this section, we will establish the stability of the balancing sequences modulo primes p congruent to $-1, -3$ modulo 8.

The following lemmas dealing with some periodicity results will prove their usefulness while proving main results of this section.

Lemma 4.1. *If $A = \{a_1, a_2, \ldots, a_r\}$ are distinct residues modulo p, then $A + mp$ for $m = 0, 1, \ldots, p^{k-1} - 1$ are also distinct residues modulo p^k.*

Proof. Suppose that for some integers $1 \le l, m \le r$ and $0 \le i, j \le p^{k-1} - 1$,

$$a_l + ip \equiv a_m + jp \pmod{p^k}. \tag{4.1}$$

This implies that $p^k \mid (i-j)p - (a_l - a_m)$ and hence, p must divide $a_l - a_m$. In other words, $a_l \equiv a_m \pmod p$, which is a contradiction since a_i's are distinct residues modulo p for $1 \le i \le r$. □

Lemma 4.2. *If $p \equiv -1 \pmod 8$, then $\pi(p) \mid \frac{p-1}{2}$. Furthermore, $\pi(p)$ is odd.*

Proof. If $p \equiv -1 \pmod 8$, then $p = 8x - 1$ for some integer x. By Theorem 2.2, $\pi(p) | p - 1 = 8x - 2$. Thus,

$$B_{8x-2} \equiv 0 \pmod p, \quad B_{8x-3} \equiv -B_1, \quad B_{8x-4} \equiv -B_2 \pmod p$$

and so on. In other words, $B_r + B_{8x-2-r} \equiv 0 \pmod p$ for $r = 1, 2, \ldots, 4x-2$. In particular, $B_{4x-2} + B_{4x} \equiv 0 \pmod p$ which implies that $6B_{4x-1} \equiv 0 \pmod p$.

Hence $B_{4x-1} = B_{\frac{p-1}{2}} \equiv 0 \pmod{p}$ as $\gcd(6,p) = 1$. We claim that $B_{\frac{p+1}{2}} \equiv 1 \pmod{p}$. Observe that

$$\mathrm{ord}_p(B_{\frac{p+1}{2}} - B_1) = \mathrm{ord}_p(2 \cdot B_{\frac{p-1}{2}} \cdot C_{\frac{p+3}{2}}) \geq 0+1+\mathrm{ord}_p\left(\frac{p-1}{2}\right)+\mathrm{ord}_p(C_{\frac{p+3}{2}}) \geq 1$$

which shows that $B_{\frac{p+1}{2}} \equiv 1 \pmod{p}$ and combining with $B_{\frac{p-1}{2}} \equiv 0 \pmod{p}$, we conclude that $\pi(p) \mid \frac{p-1}{2} = 4x-1$, which implies that $\pi(p)$ is odd. □

Lemma 4.3. *If $p \equiv -1 \pmod{8}$, then $\pi(p) = \alpha(p)$.*

Proof. By Lemma 4.2, $\pi(p)$ is odd. Thus, $B_1 + B_{\pi(p)-1} \equiv 0, B_2 + B_{\pi(p)-2} \equiv 0 \pmod{p}$, and in general $B_r + B_{\pi(p)-r} \equiv 0 \pmod{p}$ for $r = 1, 2, \ldots, \frac{\pi(p)-1}{2}$. By virtue of Theorem 2.8 of [8], $B_n | B_{kn}$ and hence if $B_n \equiv 0 \pmod{m}$, then $B_{kn} \equiv 0 \pmod{m}$. Since $B_n \equiv 0 \pmod{p}$ implies that $\alpha(p) | n$, it follows that $\alpha(p) | \pi(p)$ and thus $\alpha(p) \leq \pi(p)$. If $\alpha(p) < \pi(p)$, then $B_{\alpha(p)} \equiv 0 \pmod{p}$ implies $B_{\pi(p)-\alpha(p)} \equiv 0 \pmod{p}$. Hence, at least two B_n's out of $B_1, B_2, \ldots, B_{\pi(p)-1}$ are congruent to zero. If t be the index of the second one, then $t = 2\alpha(p)$, which shows that $2\alpha(p) < \pi(p)$ − a contradiction to $\pi(p) | 2\alpha(p)$. Therefore $\pi(p) = \alpha(p)$. □

Lemma 4.4. *If $p \equiv -3 \pmod{8}$, then $B_{\frac{p+1}{2}} \equiv 0 \pmod{p}$.*

Proof. If $p = 8x - 3$, then by Theorem 2.2,

$$B_p \equiv -1 \pmod{p}, \quad B_{p+1} \equiv 0 \pmod{p}. \tag{4.2}$$

Using the recurrence relation $B_{n+1} = 6B_n - B_{n-1}$ and (4.2) it is easy to see that

$$B_{\frac{p-1}{2}} + B_{\frac{p+3}{2}} \equiv 0 \pmod{p}. \tag{4.3}$$

Hence $6B_{\frac{p+1}{2}} \equiv 0 \pmod{p}$ and $(6,p) = 1$ implies $B_{\frac{p+1}{2}} \equiv 0 \pmod{p}$. □

Lemma 4.5. *If $p \equiv -3 \pmod{8}$, then for every x such that $0 \leq x \leq \frac{\pi(p)}{2}$,*

$$B_x \equiv B_{\frac{\pi(p)}{2}-x} \pmod{p} \quad \textit{and} \quad B_x \equiv -B_{\frac{\pi(p)}{2}+x} \pmod{p}.$$

Furthermore, $\pi(p) = 2\alpha(p)$.

Proof. If $B_x \equiv B_y \pmod{p}$ for some $0 \leq y < x \leq \frac{\pi(p)}{2}$, then $C_x \equiv \pm C_y \pmod{p}$ since $C_n = \sqrt{8B_n^2 + 1}$. Therefore, $B_{x \pm y} = B_x C_y \pm B_y C_x \equiv 0 \pmod{p}$. By Lemma 4.4, for $0 \leq x \leq \frac{\pi(p)}{2}, B_x = 0$ if and only if $x = 0, \frac{\pi(p)}{2}$. Hence $x \pm y = 0$ or $\frac{\pi(p)}{2}$. We observe that $x - y = 0$ gives trivial solution $x = y$ which is not possible since $x > y$. Again, $x + y = 0$ gives $x = -y$ which is also not possible since both x and y are non-negative and $x > y$. $x - y = \frac{\pi(p)}{2}$ gives $x = \frac{\pi(p)}{2} + y$. This is absurd since $0 \leq x \leq \frac{\pi(p)}{2}$. Thus, we are left with one

option $x+y=\frac{\pi(p)}{2}$ or equivalently, $x=\frac{\pi(p)}{2}-y$. Hence, $B_y \equiv B_{\frac{\pi(p)}{2}-y} \pmod{p}$. From 4.3 we have $B_{\frac{\pi(p)}{2}-k} \equiv -B_{\frac{\pi(p)}{2}+k} \pmod{p}$. Thus, $B_x \equiv -B_{\frac{\pi(p)}{2}+x} \pmod{p}$ and the proof is complete. □

The following two theorems, assuring the stability of the balancing sequence modulo p for $p \equiv -1, -3 \pmod 8$, are important results of this section.

Theorem 4.6. *If $p \equiv -3 \pmod 8$ and $k \in \mathbb{N}$, then $\nu_B(p^k, b) = 2$ for each feasible residue b modulo p^k. Hence the balancing sequence is stable modulo p for $p \equiv -3 \pmod 8$.*

Proof. Firstly, we will show that the number of occurrences of each feasible residue modulo p in a period is 2. In Lemma 4.5, we have shown that each feasible residue of the balancing numbers B_x, $x \in \{0, 1, \ldots, \frac{\pi(p)}{2}\}$ modulo p occurs twice. Since $B_x \equiv -B_{\frac{\pi}{2}+x} \pmod{p}$, it follows that $\#\{x : B_x \equiv b \pmod{p}, 0 \leq x \leq \pi(p)-1\} = 2$, i. e., $\nu_B(p, b) = 2$ holds for each feasible residue b modulo p. Using Lemma (4.1) we get $\nu_B(p^k, b) = 2$ for each feasible residue b modulo p^k. □

From equation (1.2), $\Omega_B(p^k) = \{2\}$.

Theorem 4.7. *If $p \equiv -1 \pmod 8$ and $k \in \mathbb{N}$, then $\nu_B(p^k, b) = 1$ for each feasible residue b modulo p^k. Hence the balancing sequence is stable modulo p for $p \equiv -1 \pmod 8$.*

Proof. Since $p \equiv -1 \pmod 8$, by Lemma 4.3 $\pi(p) = \alpha(p)$. Therefore, each feasible residue b occurs only once such that $B_r \equiv b \pmod{p}$ for $0 \leq r < \pi(p)$; otherwise $\alpha(p) < \pi(p)$. Now Lemma 4.1 confirms that $\nu_B(p^k, b) = 1$ for each feasible residue b of the balancing sequence modulo p^k. □

Therefore from equation (1.2), $\Omega_B(p^k) = \{1\}$.

5. Stability of balancing sequence modulo primes $p \equiv 1, 3 \pmod 8$

Modulo 8, there are four classes of primes $p \equiv \pm 1, \pm 3 \pmod 8$. In the last section, we have proved that the balancing sequence is stable modulo primes $p \equiv -1, -3 \pmod 8$. But unfortunately, the sequence is not stable modulo in general for primes $p \equiv 1, 3 \pmod 8$. However, for certain primes of this class, the balancing sequence is indeed stable.

The following lemmas, relating to the structure of the period and behaviour of balancing numbers occurring in a period, will play crucial roles while proving the main results of this section.

Lemma 5.1. *If $p \equiv 3 \pmod 8$, then $4 \mid \pi(p)$.*

Proof. Firstly, we will prove

$$B_{\frac{p-1}{2}} \equiv 1 \pmod p. \tag{5.1}$$

Observe that

$$\operatorname{ord}_p\Big(B_{\frac{p-1}{2}} - B_1\Big) = \operatorname{ord}_p\Big(2 \cdot B_{\frac{p-3}{4}} C_{\frac{p+1}{4}}\Big) = \operatorname{ord}_p\Big(B_{\frac{p-3}{4}}\Big) + \operatorname{ord}_p\Big(C_{\frac{p+1}{4}}\Big). \tag{5.2}$$

In view of Theorem 2.2, $\pi(p) \mid p+1$. Using this result in (5.2) we get

$$\operatorname{ord}_p\Big(B_{\frac{p-1}{2}} - B_1\Big) \geq 0 + 1 + \operatorname{ord}_p\Big(\frac{p+1}{4}\Big) \geq 1. \tag{5.3}$$

Proceeding as in Lemma 4.4 and using (4.2), it is easy to see that

$$B_{\frac{p+1}{2}} \equiv 0 \pmod p.$$

Using the recurrence $B_n = 6B_{n-1} - B_{n-2}$ and (5.1), we get $B_{\frac{p+3}{2}} \equiv 1 \pmod p$, which confirms that $\pi(p) \nmid (p+1)/2 = 4x+2$; but $\pi(p) \mid p+1 = 8x+4$ which implies that $4 \mid \pi(p)$. □

Lemma 5.2. *If $p \equiv 3 \pmod 8$ and $x \in \mathbb{N}$, then $B_{p^x \frac{\pi(p)}{4}} \not\equiv B_{3 \cdot p^x \frac{\pi(p)}{4}} \pmod p$.*

Proof. Firstly, we will show that for $x \in \mathbb{N}$

$$B_{p^x \frac{\pi(p)}{4}} \equiv (-1)^x B_{\frac{\pi(p)}{4}} \pmod p. \tag{5.4}$$

If x is even, then

$$\operatorname{ord}_p\Big(B_{p^x \frac{\pi(p)}{4}} - B_{\frac{\pi(p)}{4}}\Big) = \operatorname{ord}_p\Big(2B_{\frac{\pi(p)}{4}\frac{p^x-1}{2}} \cdot C_{\frac{\pi(p)}{4}\frac{p^x+1}{2}}\Big). \tag{5.5}$$

and $\alpha(p) \mid \frac{\pi(p)}{4} \cdot \frac{p^x-1}{2}$. Therefore, using Lemma 2.7, we get

$$\begin{aligned}
&\operatorname{ord}_p\Big(B_{p^x \frac{\pi(p)}{4}} - B_{\frac{\pi(p)}{4}}\Big) \\
&\geq \operatorname{ord}_p 2 + 1 + \operatorname{ord}_p\Big(\frac{\pi(p)}{4}\Big(\frac{p^x-1}{2}\Big)\Big) + \operatorname{ord}_p\Big(C_{\frac{\pi(p)}{4}(\frac{p^x+1}{2})}\Big) \\
&\geq 1.
\end{aligned}$$

Now, let x be odd. Since $B_{-n} = -B_n$, it can be easily proved that

$$B_{p^x \frac{\pi(p)}{4}} \equiv -B_{\frac{\pi(p)}{4}} \pmod p. \tag{5.6}$$

A similar argument as above will lead to

$$B_{3 \cdot p^x \frac{\pi(p)}{4}} \equiv (-1)^x B_{3 \cdot \frac{\pi(p)}{4}} \pmod p. \tag{5.7}$$

To complete the proof, it remains to show that $B_{\frac{\pi(p)}{4}} \not\equiv B_{3\cdot\frac{\pi(p)}{4}} \pmod p$. It is obvious since $B_{\frac{\pi(p)}{4}} \equiv -B_{3\cdot\frac{\pi(p)}{4}} \pmod p$. □

Lemma 5.3. *If $p \equiv 3 \pmod 8$ and $k \in \mathbb{N}$, then there are two distinct feasible residues of B_n with $0 \le n < \pi(p)p^{k-1}$ occurring at least $p^{[k/2]}$ times in a period modulo p^k.*

Proof. Let n be a non-negative integer, $j \in \{0,1\}$ and $\pi\big(p^{[(k-1)/2]+1}\big)|n$. We claim that

$$B_{n+\frac{\pi(p)}{4}(1+2j)p^{[(k-1)/2]}} \equiv B_{\frac{\pi(p)}{4}(1+2j)p^{[(k-1)/2]}} \pmod{p^k}. \tag{5.8}$$

If $p \equiv 3 \pmod 8$, then by virtue of Lemma 5.1, $4 \mid \pi(p)$ and $\pi(p) \mid n$ implies $4 \mid n$. Thus,

$$\begin{aligned}
&\operatorname{ord}_p\Big(B_{n+\frac{\pi(p)}{4}(1+2j)p^{[(k-1)/2]}} - B_{\frac{\pi(p)}{4}(1+2j)p^{[(k-1)/2]}}\Big)\\
&= \operatorname{ord}_p\Big(2B_{\frac{n}{2}}C_{\frac{n}{2}+\frac{\pi(p)}{4}(1+2j)p^{[(k-1)/2]}}\Big)\\
&= \operatorname{ord}_p(2) + \operatorname{ord}_p\Big(B_{\frac{n}{2}}\Big) + \operatorname{ord}_p\Big(C_{\frac{n}{2}+\frac{\pi(p)}{4}(1+2j)p^{[(k-1)/2]}}\Big)\\
&\ge 0 + 1 + \operatorname{ord}_p\Big(\frac{n}{2}\Big) + 1 + \operatorname{ord}_p\Big(\frac{n}{2} + \frac{\pi(p)}{4}(1+2j)p^{[(k-1)/2]}\Big)\\
&\ge 2(1 + [(k-1)/2]) > 2(1 + (k-1)/2 - 1) > k-1,
\end{aligned}$$

which proves (5.8). Since $\pi\big(p^{[(k-1)/2]+1}\big) \mid n$ by assumption, $\pi(p)p^{[(k-1)/2]}|n$. Therefore,

$$n = \pi(p)p^{[(k-1)/2]}i \quad \text{with some} \quad i < p^{[k/2]}. \tag{5.9}$$

Thus,

$$\begin{aligned}
0 \le n + \frac{\pi(p)}{4}(1+2j)p^{[(k-1)/2]} &= \Big(\pi(p)\cdot i + \frac{\pi(p)}{4}(1+2j)\Big)p^{[(k-1)/2]}\\
&\le \Big(\pi(p)p^{[k/2]} - \pi(p) + \frac{3\pi(p)}{4}\Big)p^{[(k-1)/2]}\\
&= \pi(p)\cdot p^{k-1} - \frac{\pi(p)}{4}p^{[(k-1)/2]} < \pi(p)\cdot p^{k-1}.
\end{aligned}$$

Now, it remains to show that $B_{\frac{\pi(p)}{4}p^{[(k-1)/2]}}$ and $B_{\frac{3\pi(p)}{4}p^{[(k-1)/2]}}$ are incongruent modulo p^k; it is enough to show that they are incongruent modulo p, which is established in Lemma 5.2. □

Lemma 5.4. *If $p \equiv 3 \pmod 8$ and $k \in \mathbb{N}$, then for every integer x with $1 \le x \le [(k-1)/2]$ there exist $(p-1)p^{k-2x-1}$ distinct feasible residue b of B_n with $0 \le n < \pi(p)p^{k-1}$ occurring at least $2p^x$ times in a period modulo p^k.*

P r o o f. Let n be a non-negative integer and $p^{x-1}\|n$. Then

$$\begin{aligned}
&\operatorname{ord}_p(B_{n+\pi(p)p^{k-x-1}} - B_n)\\
&= \operatorname{ord}_p\Big(2B_{\frac{\pi(p)}{2}p^{k-x-1}}C_{n+\frac{\pi(p)}{2}p^{k-x-1}}\Big)\\
&\geq 1 + \operatorname{ord}_p\Big(\frac{\pi(p)}{2}p^{k-x-1}\Big) + 1 + \operatorname{ord}_p\Big(n + \frac{\pi(p)}{2}p^{k-x-1}\Big)\\
&= 1 + (k-x-1) + 1 + x - 1 = k.
\end{aligned}$$

(Here we are using the inequality $\operatorname{ord}_p(a+b) \geq \min\big(\operatorname{ord}_p(a), \operatorname{ord}_p(b)\big)$ as $0 \leq x-1 \leq \frac{k-3}{2}$ and $\frac{k-1}{2} \leq k-x-1 \leq k-2$.) Therefore

$$B_{n+\pi(p)p^{k-x-1}} \equiv B_n \pmod{p^k} \tag{5.10}$$

for all n such that $p^{x-1}\|n$. We need to count the number of integers n with $0 \leq n < \pi(p)p^{k-1}$ and $p^{x-1}\|n$ for which a given b occurs as a residue of B_n modulo p^k. This is equivalent to counting the number of integers n with $0 \leq n < \pi(p)\cdot p^{k-x-1}$ and $p^{x-1}\|n$ for which a given b occurs as a residue of B_n modulo p^k and then to multiply this number by p^x. Hence we have to check the distribution of the $2(p-1)p^{k-2x-1}$ numbers

$$B_j \pmod{p^k} : 1 \leq j < \pi(p)p^{k-x-1},\ \ 2\nmid j,\ \ \frac{\pi(p)}{4} \mid j,\ \ p^{x-1}\|j. \tag{5.11}$$

We claim that half of them, i. e., $(p-1)p^{k-2x-1}$ of the B_n's are pairwise incongruent modulo p^k and other half are congruent to the first half in some way; more specifically,

$$B_{\frac{\pi(p)}{2}p^{k-x-1}-\frac{\pi(p)}{4}j} \equiv B_{\frac{\pi(p)}{4}j} \pmod{p^k},$$
$$\text{for } 1 \leq j < p^{k-x-1},\ \ 2\nmid j,\ \ p^{x-1}\|j \tag{5.12}$$

and

$$B_{\pi(p)p^{k-x-1}-\frac{\pi(p)}{4}j} \equiv B_{\frac{\pi(p)}{2}p^{k-x-1}+\frac{\pi(p)}{4}j} \pmod{p^k},$$
$$\text{for } 1 \leq j < p^{k-x-1},\ \ 2\nmid j,\ \ p^{x-1}\|j. \tag{5.13}$$

Observe that

$$\begin{aligned}
&\operatorname{ord}_p\Big(B_{\frac{\pi(p)}{2}p^{k-x-1}-\frac{\pi(p)}{4}j} - B_{\frac{\pi(p)}{4}j}\Big)\\
&= \operatorname{ord}_p\Big(2B_{\frac{\pi(p)}{4}(p^{k-x-1}-j)}C_{\frac{\pi(p)}{4}p^{k-x-1}}\Big)\\
&= \operatorname{ord}_p(2) + \operatorname{ord}_p\Big(B_{\frac{\pi(p)}{4}(p^{k-x-1}-j)}\Big) + \operatorname{ord}_p\Big(C_{\frac{\pi(p)}{4}p^{k-x-1}}\Big)\\
&\geq 1 + \operatorname{ord}_p\Big(\frac{\pi(p)}{4}(p^{k-x-1}-j)\Big) + 1 + \operatorname{ord}_p\Big(\frac{\pi(p)}{4}p^{k-x-1}\Big)\\
&= 2 + x - 1 + k - x - 1 = k
\end{aligned}$$

and

$$\begin{aligned}
&\operatorname{ord}_p\Big(B_{\pi(p)\cdot p^{k-x-1}-\frac{\pi(p)}{4}j}-B_{\frac{\pi(p)}{2}p^{k-x-1}+\frac{\pi(p)}{4}j}\Big)\\
&=\operatorname{ord}_p\Big(2B_{\frac{\pi(p)}{4}p^{k-x-1}-\frac{\pi(p)}{4}j}C_{\frac{3\pi(p)}{4}p^{k-x-1}}\Big)\\
&\geq 1+\operatorname{ord}_p\Big(\frac{\pi(p)}{4}(p^{k-x-1}-j)\Big)+1+\operatorname{ord}_p\Big(\frac{3\pi(p)}{4}p^{k-x-1}\Big)\\
&\geq 2+x-1+k-x-1=k.
\end{aligned}$$

Hence, it only remains to show that

$$B_{\pi(p)\cdot p^{k-x-1}-\frac{\pi(p)}{4}j}\not\equiv B_{\frac{\pi(p)}{4}j}\pmod{p^k},\ 1\leq j<p^{k-x-1},\ \ 2\nmid j,\ \ p^{x-1}\|j. \quad (5.14)$$

Since

$$\operatorname{ord}_p\Big(B_{\pi(p)p^{k-x-1}-\frac{\pi(p)}{4}j}-B_{-\frac{\pi(p)}{4}j}\Big)=\operatorname{ord}_p\Big(2B_{\frac{\pi(p)}{2}p^{k-x-1}}C_{\frac{\pi(p)}{2}p^{k-x-1}-\frac{\pi(p)}{4}j}\Big)\geq k,$$

we have

$$B_{\pi(p)p^{k-x-1}-\frac{\pi(p)}{4}j}\equiv B_{-\frac{\pi(p)}{4}j}\equiv -B_{\frac{\pi(p)}{4}j}\pmod{p^k}$$
$$\text{for } 1\leq j<p^{k-x-1},\ \ 2\nmid j,\ \ p^{x-1}\|j,$$

from which (5.14) follows and the proof is complete. □

Lemma 5.5. *If $p\equiv 3\pmod 8$, then there exist $\frac{\pi(p)}{2}-1$ distinct feasible residue b of B_n with $0\leq n<\pi(p)$ occurring exactly twice.*

Proof. In view of Lemma 5.3 with $k=1$, there exists two distinct feasible residue b of B_n modulo p for $n=0,1,\ldots,\pi(p)-1$ occurring only once. Hence we need to check the distribution of the remaining $\pi(p)-2$ residues, namely,

$$B_r\pmod p,\ \text{for } 0\leq r<\pi(p),\quad r\notin\left\{\frac{\pi(p)}{4},\frac{3\pi(p)}{4}\right\}.$$

We claim that half of them, i. e, $\frac{\pi(p)}{2}-1$ of B_n's are pairwise incongruent modulo p and the other half are congruent to the first half in some manner, i. e.,

$$B_i\equiv B_{\frac{\pi(p)}{2}-i}\pmod p\ \text{ and }\ B_{\frac{\pi(p)}{2}+i}\equiv B_{\pi(p)-i}\pmod p\ \text{ for }\ 0\leq i<\frac{\pi(p)}{4}. \quad (5.15)$$

But

$$\begin{aligned}
\operatorname{ord}_p(B_{\frac{\pi(p)}{2}-i}-B_i)&=\operatorname{ord}_p\Big(2B_{\frac{\pi(p)}{4}-i}C_{\frac{\pi(p)}{4}}\Big)\\
&\geq \operatorname{ord}_p(2)+1+\operatorname{ord}_p\Big(B_{\frac{\pi(p)}{4}-i}\Big)+\operatorname{ord}_p\Big(C_{\frac{\pi(p)}{4}}\Big)\geq 1
\end{aligned}$$

shows that $B_i\equiv B_{\frac{\pi(p)}{2}-i}\pmod p$. Similarly it can be easily seen that $B_{\frac{\pi(p)}{2}+i}\equiv B_{\pi(p)-i}\pmod p$. To complete the proof, it remains to show

$$B_i\not\equiv B_{\frac{\pi(p)}{2}+i}\pmod p.$$

Since, $B_i \equiv -B_{\frac{\pi(p)}{2}+i} \pmod p$ and the case $B_i \equiv 0 \equiv B_{\frac{\pi(p)}{2}+i} \pmod p$ contradicts the definition of period, it follows that $B_i \not\equiv B_{\frac{\pi(p)}{2}+i} \pmod p$. □

Remark 5.6. If $p \equiv 3 \pmod 8$ and $k \in \mathbb{N}$, then there exist $p^{k-1}(\frac{\pi(p)}{2} - 1)$ distinct feasible residues b of B_n modulo p^k with $0 \leq n < \pi(p)p^{k-1}$ occurring exactly twice.

We are now in a position to prove an important theorem of this section.

Theorem 5.7. *If $p \equiv 3 \pmod 8$ and for $i \in \{0, 1\}$, then*

$$\nu_B(p^k, b) = \begin{cases} p^{[k/2]} & \text{if } b \equiv B_{\left(\frac{\pi(p)}{4}+i\frac{\pi(p)}{2}\right)p^{[(k-1)/2]}} \pmod{p^k}, \\ 2.p^x & \text{if } b \equiv B_{\frac{\pi(p)}{4}j+i\frac{\pi(p)}{2}p^{k-x-1}}, \text{ and} \\ & \quad p^{x-1} \| j,\ 2 \nmid j,\ 1 \leq j < p^{k-x-1} \\ & \quad \text{for } x \in \{1, 2, \ldots, [(k-1)/2]\}, \\ 2 & \text{otherwise.} \end{cases}$$

P r o o f. In view of Lemma 5.3, 5.4 and Remark 5.6 we have the following results:

$$\nu_B(p^k, b) \geq p^{[k/2]},\ \nu_B(p^k, b) \geq 2 \cdot p^x \text{ and } \nu_B(p^k, b) = p^{k-1}\left(\frac{\pi(p)}{2} - 1\right). \quad (5.16)$$

Hence,

$$\begin{aligned} \sum_{b=0}^{p^k-1} \nu_B(p^k) &\geq 2p^{[k/2]} + \sum_{x=1}^{[(k-1)/2]} (p-1)p^{k-2x-1}(2p^x) + p^{k-1}\left(\frac{\pi(p)}{2} - 1\right) \\ &= \pi(p)p^{k-1}. \end{aligned} \quad (5.17)$$

In view of [9, Theorem 3.5], the left hand side of (5.17) equals $\pi(p) \cdot p^{k-1}$. Thus, equality holds in (5.16) for every feasible residue b modulo p^k. □

Remark 5.8. In the above theorem, the second case occurs if $k \geq 3$ and in this case, there are exactly $(p-1)p^{k-2x-1}$ distinct feasible residues b occur modulo p^k. In the third case, for each $k \in \mathbb{N}$, exactly $p^{k-1}\left(\frac{\pi(p)}{2}-1\right)$ distinct feasible residues b modulo p^k occur.

Using (1.2), we get $\Omega_B(p^k) = \{2, 2p, 2p^2, \ldots, 2p^{[(k-1)/2]}, p^{[k/2]}\}$. Thus, the following corollary is a direct consequence of Theorem 5.7.

Corollary 5.9. *If $p \equiv 3 \pmod 8$, then balancing sequence is not stable modulo p.*

We next search for primes $p \equiv 1 \pmod 8$ for which the balancing sequence is stable. In the following theorem, we limit the search for such primes in the class of associated Pell numbers.

Theorem 5.10. *If the prime $p \equiv 1 \pmod 8$ is an odd indexed associated Pell number, then balancing sequence is stable modulo p.*

P r o o f. Since p is an odd indexed associated Pell number, $\pi(p)$ is odd [9, Theorem 4.3]. Using the arguments given in the proof of Lemma 4.3, it is easy to see that $\pi(p) = \alpha(p)$. Now, proceeding like the proof of Theorem 4.7, one can easily verify that the balancing sequences is stable modulo such a prime. □

For some members in the class of primes $p \equiv 1 \pmod 8$, $\pi(p)$ is a multiple of 4. For example 17 is one such prime with $\pi(17) = 8$. The following theorem confirms that the balancing sequence is not stable modulo any such prime.

Theorem 5.11. *Let p be a prime such that $p \equiv 1 \pmod 8$ and $4 \mid \pi(p)$. If $i \in \{0, 1\}$, then*

$$\nu_B(p^k, b) = \begin{cases} p^{[k/2]} & \text{if } b \equiv B_{\left(\frac{\pi(p)}{4} + i \cdot \frac{\pi(p)}{2}\right) p^{[(k-1)/2]}} \pmod{p^k}, \\ 2p^x & \text{if } b \equiv B_{\frac{\pi(p)}{4} j + i \cdot \frac{\pi(p)}{2} p^{k-x-1}}, \text{ and} \\ & \quad p^{x-1} \| j,\ 2 \nmid j,\ 1 \le j < p^{k-x-1} \\ & \quad \text{for } x \in \{1, 2, \ldots, [(k-1)/2]\} \\ 2 & \text{otherwise} \end{cases}$$

P r o o f. The proof is similar to the proof of Theorem 5.7, hence it is omitted. □

There are some primes $p \equiv 1 \pmod 8$ for which $4 \nmid \pi(p)$. Such type of primes are excluded from Theorem 5.11. For example, if $p = 137, \pi(p) = 34$ and one can check that the balancing sequence is stable modulo 137. It is an open problem to identify some more subclass of primes for which the balancing sequence is stable.

Acknowledgements. It is a pleasure to thank the unanimous referee for his valuable comments and suggestions which improved the presentation of the paper to a great extent.

REFERENCES

[1] BEHERA, A.—PANDA, G. K.: *On the square roots of triangular numbers*, The Fib. Quart. **37** (1999), no. 2, 98–105.

[2] BUNDSCHUH, P.—BUNDSCHUH, R.: *The sequence of Lucas numbers is not stable modulo* 2 *and* 5, Unif. Distrib. Theory **5** (2010), 113–130.

[3] CARLIP, W.—JACOBSON, E. T.: *Unbounded stability of two-term recurrence sequences modulo* 2^k, Acta Arith. **74** (1996), 329–346.

[4] JACOBSON, E. T.: *Distribution of Fibonacci numbers mod* 2^k, The Fib. Quart. **30** (1992), 211–215.
[5] KOBLITZ, N.: *p-adic Numbers, p-adic Analysis, and Zeta-functions*. Springer, New York et al., 1984.
[6] NIEDERREITER, H.: *Distribution of Fibonacci numbers mod* 5^k, The Fib. Quart. **10** (1972), 373–374.
[7] NIVEN, I.: *Uniform distribution of sequences of integers*, Trans. Amer. Math. Soc. **98** (1961), 52–61.
[8] PANDA, G. K.: *Some fascinating properties of balancing numbers*. In: Proc Applications of Fibonacci Numbers, Congr. Numer. **194** (2006), 185–190.
[9] PANDA, G. K.—ROUT, S. S.: *Periodicity of balancing numbers*, Acta. Math. Hungar., **143** (2014), no. 2, 274–286.
[10] SOMER, L.—CARLIP, W.: *Stability of second order recurrences modulo* p^r, Int. J. Math. Math. Sci. **23** (2000), 225–241.
[11] VÉLEZ, W. Y.: *Uniform distribution of two-term recurrence sequences*, Trans. Amer. Math. Soc. **301** (1987), 37–45.

Received January 12, 2015
Accepted April 24, 2015

Sudhansu Sekhar Rout
Department of Mathematics
National Institute of Technology Rourkela
Odisha-769 008
INDIA
E-mail: sudhansumath@yahoo.com

Ravi Kumar Davala
Department of Mathematics
National Institute of Technology Rourkela
Odisha-769 008
INDIA
E-mail: davalaravikumar@gmail.com,

Gopal Krishna Panda
Department of Mathematics
National Institute of Technology Rourkela
Odisha-769 008
INDIA
E-mail: gkpanda_nit@rediffmail.com

Uniform Distribution Theory **10** (2015), no.2, 93–115

uniform distribution theory

THE h-CRITICAL NUMBER OF FINITE ABELIAN GROUPS

Béla Bajnok

Dedicated to Professor Harald Niederreiter on the occasion of his 70th birthday

ABSTRACT. For a finite abelian group G and a positive integer h, the unrestricted (resp. restricted) h-critical number $\chi(G,h)$ (resp. $\chi\hat{}(G,h)$) of G is defined to be the minimum value of m, if exists, for which the h-fold unrestricted (resp. restricted) sumset of every m-subset of G equals G itself. Here we determine $\chi(G,h)$ for all G and h; and prove several results for $\chi\hat{}(G,h)$, including the cases of any G and $h = 2$, any G and large h, and any h for the cyclic group $\mathbb{Z}_n$ of even order. We also provide a lower bound for $\chi\hat{}(\mathbb{Z}_n, 3)$ that we believe is exact for every n—this conjecture is a generalization of the one made by Gallardo, Grekos, et al. that was proved (for large n) by Lev.

Communicated by Vsevold Lev

1. Introduction

Throughout this paper, G denotes a finite abelian group of order $n \geq 2$, written in additive notation. For a positive integer h and a nonempty subset A of G, we let hA and $h\hat{}A$ denote the *h-fold unrestricted sumset* and the *h-fold restricted sumset* of A, respectively; that is, hA is the collection of sums of h not-necessarily-distinct elements of A, and $h\hat{}A$ consists of all sums of h distinct elements of A. Furthermore, we set $\Sigma A = \cup_{h=0}^{\infty} h\hat{}A$.

The study of critical numbers originated with the 1964 paper [10] of Erdős and Heilbronn, in which they asked for the least integer m so that for every set A consisting of m nonzero elements of the cyclic group $\mathbb{Z}_p$ of prime order p, we have $\Sigma A = \mathbb{Z}_p$. More generally, one can define the *critical number* of G as

$$\xi\hat{}(G) = \min\{m \ : \ A \subseteq G \setminus \{0\}, |A| \geq m \Rightarrow \Sigma A = G\}.$$

2010 Mathematics Subject Classification: 11B75.
Keywords: critical number, abelian groups, sumsets, restricted sumsets.

Note that here only subsets of $G \setminus \{0\}$ are considered; alternately, some have studied

$$\hat{\chi}(G) = \min\{m \ : \ A \subseteq G, |A| \geq m \Rightarrow \Sigma A = G\}.$$

It took nearly half a century, but now, due to the combined results of Diderrich and Mann [8], Diderrich [7], Mann and Wou [20], Dias Da Silva and Hamidoune [6], Gao and Hamidoune [14], Griggs [16], and Freeze, Gao, and Geroldinger [11, 12], we have the critical number of every group:

Theorem 1 (The combined results of authors above)**.** *Suppose that G is an abelian group of order $n \geq 10$, and let p be the smallest prime divisor of n. Then*

$$\hat{\xi}(G) = \hat{\chi}(G) - 1 = \begin{cases} \lfloor 2\sqrt{n-2} \rfloor & \text{if } G \text{ is cyclic of order } n = p \text{ or } n = pq, \\ & \text{where } q \text{ is prime and} \\ & 3 \leq p \leq q \leq p + \lfloor 2\sqrt{p-2} \rfloor + 1^1, \\ n/p + p - 2 & \text{otherwise.} \end{cases}$$

We note that, while it is easy to see that $\hat{\chi}(G)$ is at least one more than $\hat{\xi}(G)$, there is no obvious reason known for the fact that they differ by exactly one. It is also worth noting that considering unrestricted sums rather than restricted sums makes the problem trivial: the corresponding unrestricted critical numbers $\chi(G)$ and $\xi(G)$, using the notations of Theorem 1, are clearly given by

$$\xi(G) = \chi(G) - 1 = n/p.$$

We now turn to our present subject: the critical number when only a fixed number of terms are added. Here we consider both unrestricted sums and restricted sums; in particular, for a positive integer h, we define—if they exist, more on this below—the *unrestricted h-critical number* $\chi(G,h)$ and the *restricted h-critical number* $\hat{\chi}(G,h)$ as the minimum values of m for which, respectively, the h-fold sumset and the h-fold restricted sumset of every m-element subset of G is G itself:

$$\begin{aligned} \chi(G,h) &= \min\{m \ : \ A \subseteq G, |A| \geq m \Rightarrow hA = G\}, \\ \hat{\chi}(G,h) &= \min\{m \ : \ A \subseteq G, |A| \geq m \Rightarrow h\hat{A} = G\}. \end{aligned}$$

For the sake of completeness, we also discuss the two quantities:

$$\begin{aligned} \xi(G,h) &= \min\{m \ : \ A \subseteq G \setminus \{0\}, |A| \geq m \Rightarrow hA = G\}, \\ \hat{\xi}(G,h) &= \min\{m \ : \ A \subseteq G \setminus \{0\}, |A| \geq m \Rightarrow h\hat{A} = G\}. \end{aligned}$$

Let us now see when these four values exist and how the last two quantities compare to the first two. The situation for unrestricted addition is easy (see Section 2).

[1]Note that $\lfloor 2\sqrt{n-2} \rfloor = n/p + p - 1$ in this case.

Proposition 2. *Let G be an abelian group of order $n \geq 3$. Then for every $h \geq 2$, $\chi(G,h)$ and $\xi(G,h)$ exist, and $\chi(G,h) = \xi(G,h)$.*

Regarding restricted addition, for $\chi\hat{}(G,h)$ and $\xi\hat{}(G,h)$ to both exist, we clearly need $1 \leq h \leq n-1$. Furthermore, observe that if G is isomorphic to an elementary abelian 2-group, then there is no subset A of G for which $0 \in 2\hat{}A$. In Section 2 we establish the following:

Proposition 3. *Let G be an abelian group of order $n \geq 6$. Then for every $3 \leq h \leq n-3$, $\chi\hat{}(G,h)$ and $\xi\hat{}(G,h)$ exist, and $\chi\hat{}(G,h) = \xi\hat{}(G,h)$. Furthermore, the same conclusions hold if $h \in \{2, n-2\}$, unless G is isomorphic to an elementary abelian 2-group.*

According to Propositions 2 and 3, and in contrast to the situation above with an unlimited number of terms, it suffices to study $\chi(G,h)$ and $\chi\hat{}(G,h)$.

So let us see what we can say about these quantities. We can determine the exact value of $\chi(G,h)$, as follows.

Recall that the minimum size

$$\rho(G,m,h) = \min\{|hA| \ : \ A \subseteq G, |A| = m\}$$

of h-fold sumsets of m-subsets of G is known for all G, m, and h. To state the result, we need the function

$$u(n,m,h) = \min\{f_d(m,h) \ : \ d \in D(n)\},$$

where n, m, and h are positive integers, $D(n)$ is the set of positive divisors of n, and

$$f_d(m,h) = (h\lceil m/d \rceil - h + 1) \cdot d.$$

(Here $u(n,m,h)$ is a relative of the Hopf–Stiefel function used also in topology and bilinear algebra; see, for example, [24], [22], and [18].) We then have:

Theorem 4 (Plagne; cf. [23])**.** *Let n, m, and h be positive integers with $m \leq n$. For any abelian group G of order n we have*

$$\rho(G,m,h) = u(n,m,h).$$

Theorem 4 allows us to determine $\chi(G,h)$; in order to do so, we introduce a—perhaps already familiar—function first.

Suppose that h and g are fixed positive integers; since we will only need the cases when $1 \leq g \leq h$, we make that assumption here. Recall that we let $D(n)$ denote the set of positive divisors of n. We then define

$$v_g(n,h) = \max\left\{\left(\left\lfloor \frac{d-1-\gcd(d,g)}{h} \right\rfloor + 1\right) \cdot \frac{n}{d} \ : \ d \in D(n)\right\}.$$

We should note that the function $v_g(n,h)$ has appeared elsewhere in additive combinatorics already. For example, according to the classical result of Diamanda and Yap (see [5]), the maximum size of a sum-free set (that is, a set A that is disjoint from $2A$) in the cyclic group $\mathbb{Z}_n$ is given by

$$v_1(n,3) = \begin{cases} \left(1+\frac{1}{p}\right)\frac{n}{3} & \text{if } n \text{ has prime divisors congruent to 2 mod 3,} \\ & \text{and } p \text{ is the smallest such divisor,} \\ \left\lfloor \frac{n}{3} \right\rfloor & \text{otherwise;} \end{cases}$$

similarly, this author proved (see [3]) that the maximum size of a $(3,1)$-sum-free set in $\mathbb{Z}_n$ (where A is disjoint from $3A$) equals

$$v_2(n,4) = \begin{cases} \left(1+\frac{1}{p}\right)\frac{n}{4} & \text{if } n \text{ has prime divisors congruent to 3 mod 4,} \\ & \text{and } p \text{ is the smallest such divisor,} \\ \left\lfloor \frac{n}{4} \right\rfloor & \text{otherwise.} \end{cases}$$

It is believed that the analogous result for (k,l)-sum-free sets in $\mathbb{Z}_n$ (where $kA \cap lA = \emptyset$ for positive integers $k > l$) is given by $v_{k-l}(n,k+l)$; this was established for the case when $k-l$ and n are relatively prime by Hamidoune and Plagne (see [17]). In Section 3 we provide the following simpler alternate formula for $v_g(n,h)$, from which the expressions for $v_1(n,3)$ and $v_2(n,4)$ above readily follow:

Theorem 5. *Suppose that n, h, and g are positive integers and that $1 \le g \le h$. For $i = 2,3,\ldots,h-1$, let $P_i(n)$ be the set of those prime divisors of n that do not divide g and that leave a remainder of i when divided by h; that is,*

$$P_i(n) = \{\, p \in D(n) \setminus D(g) \;:\; p \text{ prime and } p \equiv i \pmod{h} \}.$$

We let I denote those values of $i = 2,3,\ldots,h-1$ for which $P_i(n) \neq \emptyset$, and for each $i \in I$, we let p_i be the smallest element of $P_i(n)$.

Then, the value of $v_g(n,h)$ is

$$v_g(n,h) = \begin{cases} \frac{n}{h} \cdot \max\left\{1 + \frac{h-i}{p_i} \;:\; i \in I\right\} & \text{if} \quad I \neq \emptyset; \\ \left\lfloor \frac{n}{h} \right\rfloor & \text{if} \quad I = \emptyset \text{ and } g \neq h; \\ \left\lfloor \frac{n-1}{h} \right\rfloor & \text{if} \quad I = \emptyset \text{ and } g = h. \end{cases}$$

Theorem 5 greatly simplifies the evaluation of the function $v_g(n,h)$.

Returning now to the h-critical number of groups, in Section 4 we prove:

Theorem 6. *For all finite abelian groups G of order n and all positive integers h, the (unrestricted) h-critical number of G equals*

$$\chi(G,h) = v_1(n,h) + 1.$$

Evaluating the restricted h-critical number $\chi^\wedge(G,h)$ seems much more challenging, and this is, of course, due to the fact that we do not have a general formula for the minimum size

$$\rho^\wedge(G,m,h) = \min\{|h^\wedge A| \ : \ A \subseteq G, |A| = m\}$$

of h-fold restricted sumsets of m-subsets of G. Indeed, we do not even know the value of $\rho^\wedge(G,m,h)$ for cyclic groups G and $h = 2$. Essentially the only general result is for groups of prime order; solving a conjecture made by Erdős and Heilbronn three decades earlier—not mentioned in [10] but in [9]—Dias Da Silva and Hamidoune succeeded in proving the following:

Theorem 7 (Dias Da Silva and Hamidoune; cf. [6])**.** *For a prime p and integers $1 \le h \le m \le p$, we have*

$$\rho^\wedge(\mathbb{Z}_p,m,h) = \min\{p, hm - h^2 + 1\}.$$

(The result was reestablished, using different methods, by Alon, Nathanson, and Ruzsa; see [1], [2], and [21].) As a consequence, we have:

Corollary 8. *For any positive integer h and prime p with $h \le p-1$ we have*

$$\chi^\wedge(\mathbb{Z}_p,h) = \lfloor (p-2)/h \rfloor + h + 1.$$

Let us see what else we can say about $\chi^\wedge(G,h)$. Trivially, for all groups G of order n we have

$$\chi^\wedge(G,1) = \chi^\wedge(G,n-1) = n.$$

In Section 5, we find the value of $\chi^\wedge(G,2)$:

Proposition 9. *Suppose that G is of order n and is not isomorphic to the elementary abelian 2-group, and let L denote its subset—indeed, subgroup—consisting of elements of order at most 2. Then*

$$\chi^\wedge(G,2) = (n + |L|)/2 + 1.$$

(Observe that $n + |L|$ is always even.) As a consequence, for high values of h, we get:

Proposition 10. *Suppose that G is of order n and is not isomorphic to the elementary abelian 2-group, and let L denote its subset consisting of elements of order at most 2. For all h with*

$$(n + |L|)/2 - 1 \le h \le n - 2,$$

we have

$$\chi^\wedge(G,h) = h + 2.$$

The easy proof is in Section 5.

Propositions 9 and 10 leave us with the task of determining $\chi\hat{}(G,h)$ for groups of composite order and

$$3 \le h \le (n+|L|)/2-2.$$

In Section 6 we complete this task for cyclic groups of even order:

Theorem 11. *Suppose that n is even and $n \ge 12$. Then*

$$\chi\hat{}(\mathbb{Z}_n,h) = \begin{cases} n/2+1 & \text{if } 3 \le h \le n/2-2; \\ n/2+2 & \text{if } h = n/2-1. \end{cases}$$

(This result was established for $h=3$ by Gallardo, Grekos, et al. in [13]; our proof for the general case is based on their method.)

In Section 7 we take a closer look at the case of $h=3$. First, we prove tight lower bounds:

Theorem 12. *Let n be an arbitrary integer with $n \ge 15$.*

(1) *If n has prime divisors congruent to 2 mod 3 and p is the smallest such divisor, then*

$$\chi\hat{}(\mathbb{Z}_n,3) \ge \begin{cases} \left(1+\frac{1}{p}\right)\frac{n}{3}+3 & \text{if } n=p; \\ \left(1+\frac{1}{p}\right)\frac{n}{3}+2 & \text{if } n=3p; \\ \left(1+\frac{1}{p}\right)\frac{n}{3}+1 & \text{otherwise.} \end{cases}$$

(2) *If n has no prime divisors congruent to 2 mod 3, then*

$$\chi\hat{}(\mathbb{Z}_n,3) \ge \begin{cases} \left\lfloor \frac{n}{3} \right\rfloor + 4 & \text{if } n \text{ is divisible by } 9; \\ \left\lfloor \frac{n}{3} \right\rfloor + 3 & \text{otherwise.} \end{cases}$$

We also claim that, actually, equality holds above for all n—this is certainly the case if n is even or prime; we have verified this (by computer) for all $n \le 50$; and in Section 7 we prove that equality follows from a conjecture that appeared in [4]. Our conjecture is a generalization of the one made by Gallardo, Grekos, et al. in [13] that was proved (for large n) by Lev in [19].

The pursuit of finding the value of $\chi\hat{}(G,h)$ in general remains challenging and exciting.

2. Preliminary results

In this section we establish Propositions 2 and 3. We start with the following easy result:

Proposition 13. *Let A be an m-subset of G and h be a positive integer.*

(1) *If either*
   (a) $h = 1$ *or*
   (b) A *is a coset of a subgroup of* G,
   then $|hA| = m$.

(2) *In all other cases,* $|hA| \geq m + 1$.

Proof. The first claim is trivial. To prove the second claim, we assume that $h \geq 2$ and that $|hA| \leq |A| = m$. We will show that for any $a \in A$, we have $A = a + H$, where H is the stabilizer subgroup of $(h-1)A$; that is,

$$H = \{g \in G \mid g + (h-1)A = (h-1)A\}.$$

Consider the set $A' = A - a$. Since $(h-1)A$ is a subset of $A' + (h-1)A$, we have

$$|hA| = |hA - a| = |A' + (h-1)A| \geq |(h-1)A| \geq |A|;$$

but then

$$A' + (h-1)A = (h-1)A.$$

Therefore, $A' \subseteq H$, and so $A \subseteq a + H$, which implies that

$$|a + H| \geq |A| \geq |hA| \geq |(h-1)A| = |H + (h-1)A| \geq |H| = |a + H|.$$

Then equality must hold throughout, and thus $a+H = A$, establishing our claim. □

As an immediate corollary, we see that $\chi(G, h)$ is well defined for all G and h, and $\xi(G, h)$ is well defined if, and only if, the trivial conditions $n \geq 3$ and $h \geq 2$ hold.

The version of Proposition 13 for restricted sumsets is substantially more complicated:

Theorem 14 (Girard, Griffiths, and Hamidoune; cf. [15])**.** *Let A be an m-subset of G, and suppose that $1 \leq h \leq m-1$. We let L denote the subgroup of G that consists of elements of order at most 2.*

(1) *If $h \in \{2, m-2\}$ and A is a coset of a subgroup of L, then $|h\hat{}A| = m - 1$.*

(2) *If any of the conditions*
   (a) $h \in \{1, m-1\}$,
   (b) A *is a coset of a subgroup of* G,

(c) $h \in \{2, m-2\}$ *and* A *consists of all but one element of a coset of a subgroup of* L*, or*
(d) $h \in \{2, m-2\}$ *and* $m = 4$ *and* A *consists of two cosets of a subgroup of order 2*

holds, then $|h\hat{}A| = m$.

(3) *In all other cases,* $|h\hat{}A| \geq m + 1$.

As a consequence, we get that $\chi\hat{}(G, h)$ is well defined if, and only if, one of the following holds:

- $h \in \{1, n-1\}$,
- $h \in \{2, n-2\}$, and G is not isomorphic to an elementary abelian 2-group,
- $3 \leq h \leq n-3$;

and $\xi\hat{}(G, h)$ is well defined if, and only if, one of the following holds:

- $n = 5$ and $h = 2$,
- $n \geq 6$, $h \in \{2, n-2\}$, and G is not isomorphic to an elementary abelian 2-group;
- $3 \leq h \leq n-3$.

From this we can conclude that, other than the trivial cases of $h \in \{1, n-1\}$ or $n \leq 5$, $\xi\hat{}(G, h)$ is well defined exactly when $\chi\hat{}(G, h)$ is.

Next we prove that our ξ quantities are equal to their respective χ versions:

Proposition 15. *When they exist, we have*

$$\xi(G, h) = \chi(G, h)$$

and

$$\xi\hat{}(G, h) = \chi\hat{}(G, h).$$

P r o o f. We only prove the first claim as the other is similar. For that, the other direction being obvious, we just need to show that

$$\xi(G, h) \geq \chi(G, h).$$

To see this, let B be a subset of G of size $\chi(G, h) - 1$ for which $hB \neq G$. Since $|B| \leq n-1$, we have $|-B| \leq n-1$ as well; let $g \in G \setminus (-B)$. Then $A = g + B$ has size $\chi(G, h) - 1$, and $A \subseteq G \setminus \{0\}$, since $0 \in A$ would contradict $g \notin -B$. But hA and hB have the same size, so we conclude that $hA \neq G$, from which our inequality follows. □

3. The function $v_g(n,h)$

In this section we prove Theorem 5. As usual, we suppose that d is a positive divisor of n, and define the function

$$f(d) = \left(\left\lfloor \frac{d-1-\gcd(d,g)}{h} \right\rfloor + 1 \right) \cdot \frac{n}{d}.$$

We first prove the following.

Claim 1. Let i be the remainder of d when divided by h. We then have

$$f(d) = \begin{cases} \frac{n}{h} \cdot \left(1 + \frac{h-i}{d}\right) & \text{if } \gcd(d,g) < i; \\ \frac{n}{h} \cdot \left(1 - \frac{h}{d}\right) & \text{if } h|d \text{ and } g = h; \\ \frac{n}{h} \cdot \left(1 - \frac{i}{d}\right) & \text{otherwise.} \end{cases}$$

Proof of Claim 1. We start with

$$\left\lfloor \frac{d-1-\gcd(d,g)}{h} \right\rfloor = \frac{d-i}{h} + \left\lfloor \frac{i-1-\gcd(d,g)}{h} \right\rfloor.$$

We investigate the maximum and minimum values of the quantity $\left\lfloor \frac{i-1-\gcd(d,g)}{h} \right\rfloor$.

For the maximum, we have

$$\left\lfloor \frac{i-1-\gcd(d,g)}{h} \right\rfloor \leq \left\lfloor \frac{(h-1)-1-1}{h} \right\rfloor \leq 0,$$

with equality if, and only if, $i-1-\gcd(d,g) \geq 0$; that is, $\gcd(d,g) < i$.

For the minimum, we get

$$\left\lfloor \frac{i-1-\gcd(d,g)}{h} \right\rfloor \geq \left\lfloor \frac{0-1-g}{h} \right\rfloor \geq \left\lfloor \frac{0-1-h}{h} \right\rfloor = -2,$$

with equality if, and only if, $i = 0$, $\gcd(d,g) = g$, and $g = h$; that is, $h|d$ and $g = h$.

The proof of Claim 1 now follows easily. □

Claim 2. Using the notations as above, assume that $\gcd(d,g) \geq i$. Then

$$f(d) \leq \begin{cases} n/h & \text{if } g \neq h; \\ (n-1)/h & \text{if } g = h. \end{cases}$$

Proof of Claim 2. By Claim 1, we have $f(d) \leq n/h$. Furthermore, unless $i = 0$ and $g \neq h$, we have

$$f(d) \leq \frac{n}{h} \cdot \left(1 - \frac{1}{d}\right) \leq \frac{n}{h} \cdot \left(1 - \frac{1}{n}\right) = \frac{n-1}{h}.$$

□

Claim 3. For all g, h, and n we have

$$v_g(n,h) \geq \begin{cases} \lfloor \frac{n}{h} \rfloor & \text{if } g \neq h; \\ \lfloor \frac{n-1}{h} \rfloor & \text{if } g = h. \end{cases}$$

Proof of Claim 3. We first note that

$$\begin{aligned} v_g(n,h) &= \max\left\{ \left(\left\lfloor \frac{d-1-\gcd(d,g)}{h} \right\rfloor + 1 \right) \cdot \frac{n}{d} \ : \ d \in D(n) \right\} \\ &\geq \left\lfloor \frac{n-1-\gcd(n,g)}{h} \right\rfloor + 1 \\ &\geq \left\lfloor \frac{n-1-g}{h} \right\rfloor + 1. \end{aligned}$$

The claim now follows, since $g+1 \leq h$, unless $g = h$ in which case

$$\left\lfloor \frac{n-1-g}{h} \right\rfloor + 1 = \left\lfloor \frac{n-1}{h} \right\rfloor.$$

□

We are now ready for the proof of Theorem 5. Let d_0 be any positive divisor of n for which $v_g(n,h) = f(d_0)$; let i_0 be the remainder of $d_0 \bmod h$. The following two claims together establish Theorem 5.

Claim 4. If $\gcd(d_0, g) \geq i_0$, then $I = \emptyset$ and

$$v_g(n,h) = \begin{cases} \lfloor \frac{n}{h} \rfloor & \text{if } g \neq h; \\ \lfloor \frac{n-1}{h} \rfloor & \text{if } g = h \end{cases}$$

Proof of Claim 4. By Claim 2,

$$v_g(n,h) = f(d_0) \leq n/h.$$

If we were to have an element $i \in I$, then for the corresponding prime divisor p_i of n we have

$$\gcd(p_i, g) = 1 < i,$$

thus by Claim 1,

$$v_g(n,h) \geq f(p_i) = \frac{n}{h} \cdot \left(1 + \frac{h-i}{p_i} \right) > \frac{n}{h},$$

a contradiction. The result now follows from Claims 2 and 3. □

Claim 5. If $\gcd(d_0, g) < i_0$, then $i_0 \in I$, $d_0 \in P_{i_0}(n)$, and

$$v_g(n,h) = \frac{n}{h} \cdot \left(1 + \frac{h - i_0}{d_0}\right).$$

P r o o f o f C l a i m 5. First, we prove that d_0 is prime. Note that our assumption implies that $i_0 \geq 2$, and thus d_0 has no divisor that is divisible by h, and has at least one prime divisor that leaves a remainder greater than 1 mod h. Let p be the smallest prime divisor of d_0 that leaves a remainder more than 1 mod h, and let i be this remainder.

We establish the inequality

$$\frac{h-2}{p^2} < \frac{h-i}{p},$$

as follows. Since $i \leq h-1$, the inequality clearly holds when $p > h-2$, so let us assume that $p \leq h-2$. Note that, in this case, $i = p$, so we need to establish that

$$\frac{h-2}{p^2} < \frac{h-p}{p};$$

this is not hard either since we have

$$h - 2 = hp - h(p-1) - 2 \leq hp - (p+2)(p-1) - 2 = hp - p^2 - p < (h-p)p.$$

Assume now that $i \neq i_0$, and thus $d_0/p \not\equiv 1 \bmod h$. Then d_0/p also has a prime divisor, say p', that leaves a remainder greater than 1 mod h, and by the choice of p, $p' \geq p$ and thus $d_0 \geq p^2$. But then we have

$$v_g(n,h) = f(d_0) = \frac{n}{h} \cdot \left(1 + \tfrac{h-i_0}{d_0}\right) \leq \tfrac{n}{h} \cdot \left(1 + \tfrac{h-2}{p^2}\right) < \tfrac{n}{h} \cdot \left(1 + \tfrac{h-i}{p}\right) = f(p),$$

a contradiction.

Therefore, $i = i_0$, and thus

$$v_g(n,h) = f(d_0) = \frac{n}{h} \cdot \left(1 + \frac{h-i_0}{d_0}\right) \leq \frac{n}{h} \cdot \left(1 + \frac{h-i_0}{p}\right) = f(p);$$

since we must have equality, $d_0 = p$ follows.

This establishes the fact that d_0 is prime. Since

$$\gcd(d_0, g) < i_0 \leq d_0,$$

d_0 cannot divide g. This establishes Claim 5, and thus completes the proof of Theorem 5. □

We should also note that it is easy to show that, when $I \neq \emptyset$ in the statement of Theorem 5, there is a unique i (and thus p_i) for which $\frac{h-i}{p_i}$ is maximal.

4. The unrestricted h-critical number

Here we establish Theorem 6; in particular, we prove that, for $m = v_1(n,h)$, we have

$$u(n,m,h) < n$$

but

$$u(n,m+1,h) \geq n.$$

Let $d_0 \in D(n)$ be such that

$$v_1(n,h) = \max\left\{\left(\left\lfloor\frac{d-2}{h}\right\rfloor+1\right)\cdot\frac{n}{d} \;:\; d\in D(n)\right\} = \left(\left\lfloor\frac{d_0-2}{h}\right\rfloor+1\right)\cdot\frac{n}{d_0}.$$

P r o o f. To establish the first inequality, simply note that

$$\begin{aligned} u(n,m,h) &\leq f_{n/d_0}(m,h) \\ &= \left(h\cdot\left(\left\lfloor\frac{d_0-2}{h}\right\rfloor+1\right)-h+1\right)\cdot\frac{n}{d_0} \\ &= \left(h\cdot\left\lfloor\frac{d_0-2}{h}\right\rfloor+1\right)\cdot\frac{n}{d_0} \\ &\leq (d_0-1)\cdot\frac{n}{d_0} \;<\; n. \end{aligned}$$

For the second inequality, we must prove that, for any $d \in D(n)$, we have $f_d(m+1,h) \geq n$; that is,

$$h\cdot\left\lceil\frac{\left(\left\lfloor\frac{d_0-2}{h}\right\rfloor+1\right)\cdot\frac{n}{d_0}+1}{d}\right\rceil - h + 1 \geq \frac{n}{d}.$$

But $n/d \in D(n)$, so by the choice of d_0, we have

$$\left(\left\lfloor\frac{d_0-2}{h}\right\rfloor+1\right)\cdot\frac{n}{d_0} \geq \left(\left\lfloor\frac{n/d-2}{h}\right\rfloor+1\right)\cdot\frac{n}{n/d},$$

and thus

$$\begin{aligned} h\cdot\left\lceil\frac{\left(\left\lfloor\frac{d_0-2}{h}\right\rfloor+1\right)\cdot\frac{n}{d_0}+1}{d}\right\rceil - h + 1 &\geq h\cdot\left\lceil\left(\left\lfloor\frac{n/d-2}{h}\right\rfloor+1\right)+\frac{1}{d}\right\rceil - h + 1 \\ &= h\cdot\left(\left\lfloor\frac{n/d-2}{h}\right\rfloor+2\right) - h + 1 \\ &\geq h\cdot\left(\frac{n/d-2-(h-1)}{h}+2\right) - h + 1 \\ &= \frac{n}{d}. \end{aligned}$$

Our proof is complete. □

5. The restricted h-critical number for $h = 2$ and large h

In this section we establish Propositions 9 and 10. We first prove the following.

Lemma 16. *For a given $g \in G$, let $L_g = \{x \in G \mid 2x = g\}$. If $L_g \neq \emptyset$, then $|L_g| = |L|$.*

Proof. Choose an element $x \in L_g$. Then $x - L_g \subseteq L$, so $|x - L_g| = |L_g| \leq |L|$.
Similarly, $x + L \subseteq L_g$, so $|x + L| = |L| \leq |L_g|$. □

Proof of Proposition 9. Suppose first that

$$m = (n + |L|)/2 + 1.$$

Note that our assumption on G implies that $3 \leq m \leq n$.

Let A be an m-subset of G, let $g \in G$ be arbitrary, and set $B = g - A$. Then $|B| = m$, and thus

$$|A \cap B| = |A| + |B| - |A \cup B| \geq 2m - n = |L| + 2.$$

By our lemma above, we must have an element $a_1 \in A \cap B$ for which $a_1 \notin L_g$. Since $a_1 \in A \cap B$, we also have an element $a_2 \in A$ for which $a_1 = g - a_2$ and thus $g = a_1 + a_2$. But $a_1 \notin L_g$, and therefore $a_2 \neq a_1$. In other words, $g \in 2\hat{}A$; since g was arbitrary, we have $G = 2\hat{}A$, as claimed.

For the other direction, we need to find a subset A of G with

$$|A| = (n + |L|)/2$$

for which $2\hat{}A \neq G$. Observe that the elements of $G \setminus L$ are distinct from their inverses, so we have a (possibly empty) subset K of $G \setminus L$ with which

$$G = L \cup K \cup (-K),$$

and L, K, and $-K$ are pairwise disjoint. Now set $A = L \cup K$. Clearly, A has the right size; furthermore, it is easy to verify that $0 \notin 2\hat{}A$ and thus $2\hat{}A \neq G$. □

Next, we show how Proposition 9 allows us evaluate $\chi\hat{}(G, h)$ for all large values of h.

Proof of Proposition 10. Assume first that A is an $(h + 1)$-subset of G. Then

$$|h\hat{}A| = h + 1 \leq n - 1,$$

so $\chi\hat{}(G, h)$ is at least $h + 2$.

Now let A be an $(h + 2)$-subset of G. Then, by symmetry, $|h\hat{}A| = |2\hat{}A|$; since

$$|A| = h + 2 \geq (n + |L|)/2 + 1,$$

by Proposition 9 we have $h\hat{}A = G$. This establishes our claim. □

6. The restricted h-critical number of cyclic groups of even order

Here we prove Theorem 11. Our methods are similar to the one by Gallardo, Grekos, et al. in [13] where they established the result for $h = 3$.

P r o o f. The cases of $h \leq 2$ or $h \geq n/2$ have been already addressed, leaving only $3 \leq h \leq n/2 - 1$. In fact, as we now show, it suffices to treat the cases of $3 \leq h \leq n/4$:

To conclude that we then have

$$\chi\hat{}(\mathbb{Z}_n, h) = n/2 + 1 \quad \text{for } \ n/4 + 1 \leq h \leq n/2 - 2$$

as well, note that, obviously, $\chi\hat{}(\mathbb{Z}_n, h) \geq n/2+1$, and that if A is a subset of $\mathbb{Z}_n$ of size $n/2 + 1$, then, since

$$3 \leq n/2 + 1 - h \leq n/4,$$

we have

$$|h\hat{}A| = |(n/2 + 1 - h)\hat{}A| = n.$$

Similarly, with $\chi\hat{}(\mathbb{Z}_n, 2) = n/2 + 2$ and $\chi\hat{}(\mathbb{Z}_n, 3) = n/2 + 1$ we can settle the case of $h = n/2 - 1$: Choosing a subset A of $\mathbb{Z}_n$ of size $n/2 + 1$ for which $|2\hat{}A| < n$ implies that we also have

$$|(n/2 - 1)\hat{}A| < n$$

and thus $\chi\hat{}(\mathbb{Z}_n, n/2-1)$ is at least $n/2+2$; while for any $B \subset \mathbb{Z}_n$ of size $n/2+2$ we get

$$|(n/2 - 1)\hat{}B| = |3\hat{}B| = n.$$

Therefore, for the rest of the proof, we assume that $3 \leq h \leq n/4$.

Since we clearly have $\chi\hat{}(\mathbb{Z}_n, h) \geq n/2 + 1$, it suffices to prove the reverse inequality. For that, let A be a subset of $\mathbb{Z}_n$ of size $n/2 + 1$; we need to prove that $h\hat{}A = \mathbb{Z}_n$.

Let O and E denote the set of odd and even elements of $\mathbb{Z}_n$, respectively, and let A_O and A_E be the set of odd and even elements of A, respectively. Note that both A_O and A_E have size at most $n/2$ and thus neither can be empty. We will consider four cases:

Assume first that $|A_O| \leq 2$. Then $|A_E| \geq n/2 - 1$. Observe that $3 \leq h \leq n/4$ and $n \geq 12$ imply that

$$2 \leq h - 1 < h \leq n/2 - 3,$$

and $n/2 - 1$ is not a divisor of n. Therefore, by Theorem 14, both $(h - 1)\hat{}A_E$ and $h\hat{}A_E$ have size at least $n/2$. But, of course, both $(h - 1)\hat{}A_E$ and $h\hat{}A_E$ are subsets of E, so

$$(h - 1)\hat{}A_E = h\hat{}A_E = E.$$

Now let a be any element of A_O; we then see that

$$a + (h-1)\hat{}A_E = a + E = O.$$

Therefore,

$$(a + (h-1)\hat{}A_E) \cup h\hat{}A_E = O \cup E = \mathbb{Z}_n;$$

since both $a + (h-1)\hat{}A_E$ and $h\hat{}A_E$ are subsets of $h\hat{}A$, we get $h\hat{}A = \mathbb{Z}_n$.

Next, we assume that $|A_E| \leq 2$. In this case, an argument similar to the one in the previous case yields that

$$(h-1)\hat{}A_O = \begin{cases} O & \text{if } h \text{ is even,} \\ E & \text{if } h \text{ is odd;} \end{cases}$$

and

$$h\hat{}A_O = \begin{cases} E & \text{if } h \text{ is even,} \\ O & \text{if } h \text{ is odd.} \end{cases}$$

Let a be any element of A_E; we get

$$(a + (h-1)\hat{}A_O) \cup h\hat{}A_O = \mathbb{Z}_n$$

regardless of whether h is even or odd; therefore, $h\hat{}A = \mathbb{Z}_n$.

Before turning to the last two cases, we observe that, since $h \leq n/4$, we have

$$|A| = n/2 + 1 \geq 2h + 1,$$

and thus at least one of A_O or A_E must have size at least $h+1$.

Consider the case when $|A_O| \geq 3$ and $|A_E| \geq h+1$. Referring to Theorem 14 again, we deduce that $(h-2)\hat{}A_E$ and $(h-1)\hat{}A_E$ both have size at least $|A_E|$, and that $2\hat{}A_O$ is of size at least $|A_O|$.

Now let g_O be any element of O; we have

$$|g_O - A_O| + |(h-1)\hat{}A_E| \geq |A_O| + |A_E| = n/2 + 1.$$

But $g_O - A_O$ and $(h-1)\hat{}A_E$ are both subsets of E, so they cannot be disjoint; this then means that g_O can be written as the sum of an element of A_O and $h-1$ distinct elements of A_E, so $g_O \in h\hat{}A$.

Similarly, for any element g_E of E, we have

$$|g_E - (h-2)\hat{}A_E| + |2\hat{}A_O| \geq |A_E| + |A_O| = n/2 + 1,$$

and thus g_E can be written as the sum of $h-2$ distinct elements of A_E and two distinct elements of A_O, so $g_E \in h\hat{}A$.

Combining the last two paragraphs yields $O \cup E \subseteq h\hat{}A$ and thus $h\hat{}A = \mathbb{Z}_n$.

For our fourth case, assume that $|A_E| \geq 3$ and $|A_O| \geq h+1$. As above, we can conclude that $|(h-2)\hat{}A_O| \geq |A_O|$, $|(h-1)\hat{}A_O| \geq |A_O|$, and $|2\hat{}A_E| \geq |A_E|$.

Let g be any element of $\mathbb{Z}_n$. If g and h are of the same parity (both even or both odd), then we find that $g-(h-2)\hat{}A_O$ and $2\hat{}A_E$ are each subsets of E. As above, we see that they cannot be disjoint, and thus

$$g \in (h-2)\hat{}A_O + 2\hat{}A_E \subseteq h\hat{}A.$$

The subcase when g is even and h is odd is similar: this time we see that $g-(h-1)\hat{}A_O$ and A_E are each subsets of E and that they cannot be disjoint, so

$$g \in (h-1)\hat{}A_O + A_E \subseteq h\hat{}A.$$

The final subcase, when g is odd and h is even, needs more work. We first prove that there is at most one element $a \in A_O$ for which $A_O \setminus \{a\}$ is the coset of a subgroup of $\mathbb{Z}_n$. Suppose, indirectly, that a_1 and a_2 are distinct elements of A_O so that $A_O \setminus \{a_1\}$ and $A_O \setminus \{a_2\}$ are both cosets. In this case, they must be cosets of the same subgroup since $\mathbb{Z}_n$ has only one subgroup of that size. But $|A_O| \geq 3$, so $A_O \setminus \{a_1\}$ and $A_O \setminus \{a_2\}$ are not disjoint, which implies that they are actually equal, which is a contradiction since a_1 is an element of $A_O \setminus \{a_2\}$ but not of $A_O \setminus \{a_1\}$.

We also need to consider the special case when $|A_O| = 5$; we can then see that there is at most one element $a \in A_O$ for which $A_O \setminus \{a\}$ is the union of two cosets of $\{0, n/2\}$.

Hence we have an element $a_O \in A_O$ so that $A_O \setminus \{a_O\}$ is not the coset of a subgroup of $\mathbb{Z}_n$, and not the union of two cosets of the subgroup of size 2. But then, by Theorem 14,

$$|(h-2)\hat{}(A_O \setminus \{a_O\})| \geq |A_O|.$$

Therefore,

$$|(h-2)\hat{}(A_O \setminus \{a_O\})| + |g - a_O - A_E| \geq |A_O| + |A_E| = n/2 + 1;$$

since both

$$(h-2)\hat{}(A_O \setminus \{a_O\}) \quad \text{and} \quad g - a_O - A_E$$

are subsets of E, this can only happen if they are not disjoint, which means that

$$g \in (h-2)\hat{}(A_O \setminus \{a_O\}) + (a_O + A_E) \subseteq h\hat{}A.$$

This completes our proof. □

7. The restricted 3-critical number of cyclic groups

In this section we summarize what we can say about the case of $h = 3$ in the cyclic group of order n. Recall that by Theorem 4, we have

$$\rho(\mathbb{Z}_n, m, h) = u(n, m, h) = \min\left\{(h\lceil m/d\rceil - h + 1)\cdot d \mid d \in D(n)\right\}.$$

We will rely on the following result on the minimum size of h-fold restricted sumsets:

Theorem 17 (B.; cf. [4]). *Suppose that positive integers n and m satisfy $4 \le m \le n$, and let $u_3 = u(n, m, 3)$ and $d_0 = \gcd(n, m-1)$. We then have:*

$$\hat{\rho}(\mathbb{Z}_n, m, 3) \le \begin{cases} \min\{u_3, 3m - 3 - d_0\} & \text{if } d_0 \ge 8; \\ \min\{u_3, 3m - 10\} & \text{if } d_0 = 7, \text{ or} \\ & d_0 \le 5,\ 3|n, \text{ and } 3|m, \text{ or} \\ & d_0 \le 5,\ (3m-9)|n, \text{ and } 5|(m-3); \\ \min\{u_3, 3m - 9\} & \text{if } d_0 = 6, \text{ or} \\ & m = 6 \text{ and } 10|n \text{ but } 3 \not| n; \\ \min\{u_3), 3m - 8\} & \text{otherwise.} \end{cases}$$

Proof of Theorem 12. Note that the case when n is even follows from Theorem 11, since

$$\left(1 + \frac{1}{2}\right)\frac{n}{3} + 1 = \frac{n}{2} + 1;$$

and the case when n is prime follows from Corollary 8 since

$$\left\lfloor \frac{p-2}{3} \right\rfloor + 3 + 1 = \begin{cases} \left(1 + \frac{1}{p}\right)\frac{p}{3} + 3 & \text{if } p \equiv 2 \bmod 3; \\ \lfloor \frac{p}{3} \rfloor + 3 & \text{otherwise.} \end{cases}$$

Therefore, we may assume that n is odd and composite, and $n \ge 21$.

We observe first that for

$$m = \left\lfloor \frac{n}{3} \right\rfloor + 2$$

we have

$$\hat{\rho}(\mathbb{Z}_n, m, 3) \le \hat{u}(n, m, 3) \le 3m - 8 \le n - 2,$$

so we always have

$$\hat{\chi}(\mathbb{Z}_n, 3) \ge \left\lfloor \frac{n}{3} \right\rfloor + 3.$$

Assume now that n has no prime divisors congruent to 2 mod 3 and that n is divisible by 9; let $m = n/3 + 3$. Then $m - 1$ and n are relatively prime, since if d is a divisor of both $m - 1$ and n, then d will divide both $3m - 3$ and n,

and hence also their difference, which is 6. However, n is odd and $m-1$ is not divisible by 3 (since m is), so $d=1$. According to Theorem 17,

$$\hat{\rho}(\mathbb{Z}_n, m, 3) \leq \min\{u(n,m,3), 3m-10\} \leq 3m-10 = n-1,$$

so

$$\hat{\chi}(\mathbb{Z}_n, 3) \geq n/3+4.$$

Suppose now that n has a prime divisors congruent to 2 mod 3, and let p be the smallest of these. We then have

$$\hat{\chi}(\mathbb{Z}_n, 3) \geq \chi(\mathbb{Z}_n, 3) = v_1(n,3)+1 = \left(1+\frac{1}{p}\right)\frac{n}{3}+1.$$

Now if $n=3p$, then we further have

$$\hat{\chi}(\mathbb{Z}_n, 3) \geq \left(1+\frac{1}{p}\right)\frac{n}{3}+2,$$

since for

$$m = \left(1+\frac{1}{p}\right)\frac{n}{3}+1 = p+2$$

we have

$$\hat{\rho}(\mathbb{Z}_n, m, 3) \leq \hat{u}(n,m,3) \leq 3m-8 = 3p-2 = n-2.$$

Our proof is now complete. □

In [4] we made the following conjecture:

Conjecture 18. *For all n and m with $4 \leq m \leq n$, we have equality in Theorem 17.*

Correspondingly, we believe that:

Conjecture 19. *For all values of $n \geq 15$, equality holds in Theorem 12.*

We have verified that Conjecture 19 holds for all values of $n \leq 50$, and by Corollary 8 and Theorem 11, it holds when n is prime or even. As additional support, we prove the following:

Theorem 20. *Conjecture 18 implies Conjecture 19.*

Proof. As we noted before, we may assume that n is odd, composite, and greater than 15.

Suppose first that n has a prime divisor that is congruent to 2 mod 3, and let p be the smallest such prime; since n is odd, $p \geq 5$. Let us set

$$m = \left(1+\frac{1}{p}\right)\frac{n}{3}+1.$$

We need to prove that Conjecture 18 implies both of the following statements:

A: $\hat{\rho}(\mathbb{Z}_n, m+1, 3) = n$.

B: If $\hat{\rho}(\mathbb{Z}_n, m, 3) < n$, then $n = 3p$.

First, note that $m = \chi(\mathbb{Z}_n, 3)$, so $u(n, m, 3) = n$ and thus $u(n, m+1, 3) = n$ as well. Thus, looking at the conjectured formula for $\hat{\rho}(\mathbb{Z}_n, m, 3)$, to prove statement A, it suffices to verify that

A.1: $3(m+1) - 3 - \gcd(n, (m+1) - 1) \geq n$;

A.2: $3(m+1) - 9 \geq n$; and

A.3: If $3(m+1) - 10 < n$, then $\gcd(n, (m+1) - 1) \neq 7$, $m+1$ is not divisible by 3, and $(m+1) - 3$ is not divisible by 5.

Observe that if d divides both n and m, then d divides $3m - n$ as well, and so

$$\gcd(n, m) \leq 3m - n = n/p + 3,$$

which implies that

$$3(m+1) - 3 - \gcd(n, (m+1) - 1) \geq (p+1) \cdot n/p + 3 - (n/p + 3) = n,$$

proving A.1.

To prove A.2, observe that, since n is neither prime nor even, we have $n \geq 3p$, and so

$$3(m+1) - 9 = (p+1) \cdot n/p - 3 \geq n.$$

Similarly, we see that $3(m+1) - 10 < n$ may only occur if $n = 3p$, in which case $m = p + 2$, but then neither 3 nor p divides m, so $\gcd(n, m) = 1$; $m + 1 = p + 3$ is not divisible by 3; furthermore, $m - 2 = p$ is not divisible by 5 (since $p = 5$ would give $n = 15$, which we excluded). This proves A.3.

To prove statement B, we will suppose, indirectly, that $n \neq 3p$. But we assumed that n was odd and composite, so $n = 5p$ or $n \geq 7p$; furthermore, if $n = 5p$ then, for p to be the smallest prime divisor of n that is congruent to 2 mod 3, p would need to be 5. For $n = 25$ we get $m = 11$, but Conjecture 18 implies that $\hat{\rho}(\mathbb{Z}_{25}, 11, 3) = 25$, so we can rule out $n = 25$ and so assume that $n \geq 7p$. Thus, looking again at the conjectured formula for $\hat{\rho}(\mathbb{Z}_n, m, 3)$, to prove statement B, it suffices to verify that

B.1: $3m - 3 - \gcd(n, m - 1) \geq n$; and

B.2: If $n \geq 7p$, then $3m - 10 \geq n$.

The proofs of B.1 and B.2 are similar to that of A.1 and A.2, respectively—we omit the details. This completes the proof of statement B.

Assume now that n has no prime divisors congruent to 2 mod 3. This, of course, means that n itself is not congruent to 2 mod 3. We set

$$m = \left\lfloor \frac{n}{3} \right\rfloor + 3.$$

We need to prove that Conjecture 18 implies both of the following statements:

C: $\hat{\rho}(\mathbb{Z}_n, m+1, 3) = n$.

D: If $\hat{\rho}(\mathbb{Z}_n, m, 3) < n$, then n is divisible by 9.

This time we have $m = \chi(\mathbb{Z}_n, 3)+2$, so $u(n, m, 3) = n$ and thus $u(n, m+1, 3) = n$ as well. Thus, looking at the conjectured formula for $\hat{\rho}(\mathbb{Z}_n, m, 3)$, to prove statement C, it suffices to verify that

C.1: $3(m+1) - 3 - \gcd(n, (m+1)-1) \geq n$;

C.2: $3(m+1) - 10 \geq n$.

Suppose that d divides both n and m, then d divides

$$3m - n = \begin{cases} 9 & \text{if } n \equiv 0 \bmod 3; \\ 8 & \text{if } n \equiv 1 \bmod 3. \end{cases}$$

Therefore,

$$3(m+1) - 3 - \gcd(n, (m+1)-1) \geq \begin{cases} n + 12 - 3 - 9 & \text{if } n \equiv 0 \bmod 3; \\ n - 1 + 12 - 3 - 8 & \text{if } n \equiv 1 \bmod 3. \end{cases}$$

This proves C.1. Since

$$m + 1 \geq (n-1)/3 + 4,$$

statement C.2 follows as well.

To prove statement D, we first prove that $\gcd(n, m-1) \leq 5$. Indeed, if d is a divisor of both n and $m-1$, then d divides $3m - 3 - n$, which is at most 6; however d cannot be 6 as n is odd. We also see that

$$3m - 8 \geq n - 1 + 9 - 8 = n.$$

Furthermore, $m \neq 6$ since $n > 15$.

Therefore, according to Conjecture 18, for $\hat{\rho}(\mathbb{Z}_n, m, 3)$ to be less than n, we must have either n and m both divisible by 3, or n divisible by $3m - 9$ and $m - 3$ divisible by 5. Since in both these cases n is divisible by 3, we have $m = n/3 + 3$. We can rule out the second possibility: if $m - 3 = n/3$ were to be divisible by 5, then n would be as well, contradicting our assumption that n has no prime divisors congruent to 2 mod 3. This leaves only one possibility: that n and m are both divisible by 3, which implies that n is divisible by 9, as claimed. Our proof of statement D and thus of Theorem 20 is now complete. □

It is worth mentioning that, as a special case of Conjecture 19, for odd integers $n \geq 31$,

$$\hat{\chi}(\mathbb{Z}_n, 3) \leq \tfrac{2}{5}n + 1.$$

(The additive constant could be adjusted to include odd integers less than 31.) This conjecture was made by Gallardo, Grekos, et al. in [13], and (for large n) proved by Lev via the following more general result:

Theorem 21 (Lev; cf. [19]). *Let G be an abelian group of order n with*

$$n \geq 312|L| + 923,$$

where, as before, L is the collection of elements of G that have order at most 2. Then for any subset A of G, at least one of the following possibilities holds:

- $|A| \leq \frac{5}{13}n$;
- *A is contained in a coset of an index-two subgroup of G;*
- *A is contained in a union of two cosets of an index-five subgroup of G; or*
- $3\hat{}A = G$.

So, in particular, if n is odd, is at least 1235, and a subset A of $\mathbb{Z}_n$ has size more than $2n/5$, then the last possibility must hold, so we get:

Corollary 22. *If $n \geq 1235$ is an odd integer, then*

$$\chi\hat{}(\mathbb{Z}_n, 3) \leq \tfrac{2}{5}n + 1.$$

The bound on n in Corollary 22 can hopefully be reduced.

As another special case of Conjecture 19, we claim that if $n \geq 83$ is odd and not divisible by five, then

$$\chi\hat{}(\mathbb{Z}_n, 3) \leq \tfrac{4}{11}n + 1.$$

Theorem 21 does not quite yield this: while a careful read of [19] enables us to reduce the coefficient $5/13$ to $(3 - \sqrt{5})/2$ (at least for large enough n), this is still higher than $4/11$.

It is also worth pointing out that combining Theorem 6 with Conjecture 19 yields that, when $n \geq 15$, we have

$$\chi(\mathbb{Z}_n, 3) \leq \chi\hat{}(\mathbb{Z}_n, 3) \leq \chi(\mathbb{Z}_n, 3) + 3.$$

This is in contrast to the fact that for every positive integer C, there are values of n and m so that the quantities $\rho\hat{}(\mathbb{Z}_n, m, 3)$ and $\rho(\mathbb{Z}_n, m, 3)$ are further than C away from one another (cf. [4]).

Acknowledgements. The author acknowledges preliminary work by J. Butterworth and K. Campbell on Theorems 5 and 6, respectively.

REFERENCES

[1] ALON, N.—NATHANSON, M. B.—RUZSA, I.: *Adding Distinct Congruence Classes Modulo a Prime*, Amer. Math. Monthly, **102** (1995), 250–255.

[2] ALON, N.—NATHANSON, M. B.—RUZSA, I.: *The Polynomial Method and Restricted Sums of Congruence Classes*, J. Number Theory **56** (1996), 404–417.

[3] BAJNOK, B.: *On the maximum size of a (k, l)-sum-free subset of an abelian group*, Int. J. Number Theory **5**(6) (2009), 953-971.

[4] BAJNOK, B.: On the minimum size of restricted sumsets in cyclic groups, (to appear); `www.arxiv.org/pdf/1305.2141`

[5] DIANANDA, P. H.—YAP, H. P.: *Maximal sum-free sets of elements of finite groups*, Proceedings of the Japan Academy **45** (1969), 1–5.

[6] DIAS DA SILVA, J. A. — Y. O. HAMIDOUNE, Y. O.: *Cyclic space for Grassmann derivatives and additive theory*. Bull. London Math. Soc. **26** (1994), 140–146.

[7] DIDERRICH, G. T.: *An Addition Theorem for Abelian Groups of Order pq*, J. Number Theory **7** (1975), 33–48.

[8] DIDERRICH, G. T.—MANN, H. B.: *Combinatorial Problems in Finite Abelian Groups*. In: A Survey of Combinatorial Theory (J. N. Srivastava et al., eds.), North-Holland, 1973.

[9] ERDŐS, P.—GRAHAM, R. L.: *Old and New Problems and Results in Combinatorial Number Theory*. L'Enseignement Mathématique, Geneva, 1980.

[10] ERDŐS, P.—HEILBRONN, H.: *On the addition of residue classes* (mod p), Acta Arith. **9** (1964), 149–159.

[11] FREEZE, M.—GAO, W.—GEROLDINGER, A.: *The critical number of finite abelian groups*, J. Number Theory **129** (2009), 2766–2777.

[12] FREEZE, M.—GAO, W.—GEROLDINGER, A.: *Corrigendum to "The critical number of finite abelian groups, J. Number Theory* **129** *(2009), 2766–2777"* (submitted to J. Number Theory).

[13] GALLARDO, L.—GREKOS, G., ET AL.,: *Restricted addition in $\mathbb{Z}/n\mathbb{Z}$ and an application to the Erdős-Ginzburg-Ziv problem* J. London Math. Soc. (2) **65** (2002), 513–523.

[14] GAO, W. —HAMIDOUNE, Y. O.: *On additive bases*, Acta Arithmetica, **88** (1999), no. 3, 233-237.

[15] GIRARD, B.—GRIFFITHS, S.—HAMIDOUNE, Y. O.: *k-sums in abelian groups*, Combin. Probab. Comput. **21** (2012), no. 4. 582–596.

[16] GRIGGS, J. R.: *Spanning subset sums for Finite Abelian groups*, Discrete Mathematics **229** (2001), 89–99.

[17] HAMIDOUNE, Y. O.—PLAGNE, A.: A new critical pair theorem applied to sum-free sets in Abelian groups, Comment. Math. Helv. **XX** (2003), 1–25.

[18] KÁROLYI, GY.: A note on the Hopf–Stiefel function. European J. Combin. **27** (2006), 1135–1137.

[19] LEV, V.: Three-fold Restricted Set Addition in Groups, European J. Combin. **23** (2002), 613–617.

[20] MANN, H. B.—WOU, Y. F.: *Addition theorem for the elementary abelian group of type (p, p)*, Monatshefte für Math. **102** (1986), 273–308.

[21] NATHANSON, M.: *Additive Number Theory: Inverse Problems and the Geometry of Sumsets*. Graduate Texts in Mathematics Vol. 165, Springer-Verlag 1996.

[22] PLAGNE, A.: *Additive number theory sheds extra light on the Hopf-Stiefel ∘ function*, Enseign. Math., II Sér. **49** (2003), no. 1–2 109–116.

[23] PLAGNE, A.: *Optimally small sumsets in groups, I. The supersmall sumset property, the $\mu_G^{(k)}$ and the $\nu_G^{(k)}$ functions*, Unif. Distrib. Theory **1** (2006), no. 1, 27–44.

[24] SHAPIRO, D.: *Products of sums of squares*, Expo. Math. **2** (1984), 235–261.

Received December 7, 2014
Accepted May 16, 2015

Béla Bajnok
Department of Mathematics
Gettysburg College
300 N. Washington Street
Gettysburg, PA 17325
U.S.A.
E-mail: bbajnok@gettysburg.edu

Uniform Distribution Theory **10** (2015), no.2, 117–183

uniform distribution theory

SOME APPLICATIONS OF DISTRIBUTION FUNCTIONS OF SEQUENCES

Oto Strauch

Dedicated to Prof. H. Niederreiter on the occasion of his 70th birthday

ABSTRACT. This expository paper presents some old and some new results on distribution functions of sequences $x_n \in [0,1)$, $n = 1, 2, \ldots$ Firstly we describe old applications: Statistically independent sequences; statistically convergent sequences; statistical limit points; and uniform maldistributed sequences. Then we give some recent results: Benford's law; copulas; and ratio sequences. Secondly we present some methods for computing the set $G(x_n)$ of all distribution functions of x_n: directly by definition of distribution functions; using connectivity of $G(x_n)$; solving a moment problem $X_1 = \int_0^1 g(x)\mathrm{d}x$, $X_2 = \int_0^1 xg(x)\mathrm{d}x$ and $X_3 = \int_0^1 g^2(x)\mathrm{d}x$ for distribution functions $g(x)$; and mapping x_n to $f(x_n)$, for some function $f : [0,1] \to [0,1]$. Parts of this paper were presented at the UDT conferences in Marseilles 2008, Strobl 2010, Smolenice 2012, and Ostravice 2014 and also in MCQMC conference, Warszawa 2010.

Communicated by Werner Georg Nowak

Contents

2010 Mathematics Subject Classification: 11K06, 11K31.
Keywords: Distribution function, copula, statistical independent sequences, statistical limit, sequence of logarithm, moment problem, Benford's law, prime number, moment problem, Riemann-Stieltjes integral, von Neumann-Kakutani transformation.
Supported by the VEGA Project 2/0146/14.

1. Introduction

Let $\mathbf{x}_n$, $n = 1, 2, \ldots$, be an infinite sequence in the s-dimensional unit cube $[0,1)^s$. Denote the step distribution function $F_N(\mathbf{x})$ of $\mathbf{x}_1, \ldots, \mathbf{x}_N$ as

$$F_N(\mathbf{x}) = \frac{\#\{n \le N; \mathbf{x}_n \in [\mathbf{0}, \mathbf{x})\}}{N}, \tag{1}$$

where $[\mathbf{0}, \mathbf{x}) = [0, x_1) \times \cdots \times [0, x_s)$. By Riemann-Stieltjes integration we have

$$\frac{1}{N_k} \sum_{n=1}^{N_k} f(\mathbf{x}_n) = \int_{[\mathbf{0},\mathbf{1}]} f(\mathbf{x})\mathrm{d}F_{N_k}(\mathbf{x}). \tag{2}$$

By Helly theorem, for continuous $f(\mathbf{x})$ and for a weak limit

$$\lim_{k\to\infty} F_{N_k}(\mathbf{x}) \ = \ g(\mathbf{x}) \tag{3}$$

we have

$$\lim_{k\to\infty} \frac{1}{N_k} \sum_{n=1}^{N_k} f(\mathbf{x}_n) \Longrightarrow \int_{[\mathbf{0},\mathbf{1}]} f(\mathbf{x})\mathrm{d}g(\mathbf{x}). \tag{4}$$

The function $g(\mathbf{x})$ in (3) is called distribution function (abbreviating d.f.) of $\mathbf{x}_n$ [1] Denote by $G(\mathbf{x}_n)$ the set of all possible limits in (3), for an arbitrary $N_1 < N_2 < \dots$ and the given infinite sequence $\mathbf{x}_n$, $n = 1, 2, \dots$

This expository paper is devoted specially, for employing and calculating $G(\mathbf{x}_n)$ in the dimension $s = 1$ and $s = 2$.

The study of the set of d.f.s of a sequence, still unsatisfactory today, was initiated by J. G. van der Corput [46]. The one-element set $G(\mathbf{x}_n) = \{g(\mathbf{x})\}$ correspond to the notion of asymptotic distribution function (abbreviating a.d.f.) $g(\mathbf{x})$. In the case $g(\mathbf{x}) = \mathbf{x}$ the sequence is called uniformly distributed (abbreviating u.d.) In the monograph L. Kuipers and H. Niederreiter [22] to d.f.s is devoted Chapter 7, pp. 53–68, and in M. Drmota and R.F. Tichy [9] Part 1.5, pp. 138–153. [2] The a.d.f. of a sequence x_n was introduced by I.J. Schoenberg [34]. Main goal of this paper is a propagation of some partial results in the theory of d.f.s.

The outline of our paper is as follows.

In Section 2 we characterize some known classes of sequences x_n, originally defined by some properties of x_n, by using the set $G(x_n)$ of all distribution functions of x_n :

- Statistically independent sequences.
- Statistically convergent sequences.
- Statistical limit points.
- Uniform maldistributed sequences.

Then we give some recent results:

- Benford's law.
- Copulas.
- Ratio sequences.

In Section 3 we present some methods for computing $G(x_n)$, namely:

- Directly by definition of d.f.s.
- Using connectivity of $G(x_n)$.
- Solving moment problem $(X_1, X_2, X_3) = \left(\int_0^1 g(x)\mathrm{d}x, \int_0^1 xg(x)\mathrm{d}x, \int_0^1 g^2(x)\mathrm{d}x \right)$.

[1] The limit (4) generalizes $\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{\infty} f(x_n) = \int_0^1 f\mathrm{d}x$ - the fundamental Weyl's limit relation holding for any continuous function $f(x)$ defined on $[0, 1]$ and any uniformly distributed sequence x_n.

[2] Some authors, see R. Winkler [50], instead distribution functions $g(x)$ use measures μ induced by the interval (x, y) measure $\mu((x, y)) = g(y) - g(x)$.

• Mapping x_n to $f(x_n)$.

To clarify methods we add, in some places, sketch of proofs and examples. [3] [4]

2. Examples of applications of $G(x_n)$

We repeat definitions in the Introduction for dimension $s = 1$ following monographs [22], [25], [9] and [44]:

• a sequence x_n, $n = 1, 2, \ldots$, $x_n \in [0, 1)$.

• Define step d.f. of x_n as

$$F_N(x) = \frac{\#\{n \le N; x_n \in [0, x)\}}{N}.$$

• A function $g : [0, 1] \to [0, 1]$ is d.f. of x_n if there exists a sequence of indices $N_1 < N_2 < \ldots$ such that $F_{N_k}(x) \to g(x)$ for all continuity points x of $g(x)$ as $k \to \infty$.

• The set of all such $g(x)$ we shall denote by $G(x_n)$ and the notion of the distribution of x_n we shall identify with $G(x_n)$, i.e., the distribution of x_n is known if we known the set $G(x_n)$.

2.1. Basic properties of $G(x_n)$

For every sequence $x_n \in [0, 1)$:

• $G(x_n)$ is non–empty, and it is either a singleton or has infinitely many elements.

• $G(x_n)$ is closed and connected in the topology of the weak convergence defined by the metric

$$d(g_1, g_2) = \sqrt{\int_0^1 \big(g_1(x) - g_2(x)\big)^2 \mathrm{d}x}. \tag{5}$$

These properties are characteristic for a set of d.f.s:

• Given a non–empty set H of distribution functions, there exists a sequence x_n in $[0, 1)$ such that $G(x_n) = H$ if and only if H is closed and connected.

• First Helly theorem (or Helly selection principle): Any sequence $g_n(x)$ of d.f. contains a subsequence $g_{k_n}(x)$ such that the sequence $g_{k_n}(x)$ converges for every $x \in [0, 1]$ and its point limit $\lim_{n\to\infty} g_{k_n}(x) = g(x)$ is also a d.f.

• Second Helly theorem (or Helly-Bray theorem): If we have $\lim_{n\to\infty} g_n(x) = g(x)$ a.e. on $[0, 1]$, then for a continuous function $f : [0, 1] \to \mathbb{R}$ we have $\lim_{n\to\infty} \int_0^1 f(x)\mathrm{d}g_n(x) = \int_0^1 f(x)\mathrm{d}g(x)$.

[3] In each paragraph we shall numbering figures starting from 1.

[4] We shall see in many cases that we need solve corresponding functional equations.

• The upper $\underline{g}(x)$ and the lower $\overline{g}(x)$ d.f.s are
$\liminf_{N\to\infty} F_N(x) = \underline{g}(x), \limsup_{N\to\infty} F_N(x) = \overline{g}(x)$.
It is equivalent to
$\underline{g}(x) = \inf_{g\in G(x_n)} g(x), \overline{g}(x) = \sup_{g\in G(x_n)} g(x)$.

A connectivity of $G(x_n)$ can be proved by the following theorem of Barone [4]

Theorem 1 (H.G. Barone (1939)). *If t_n, $n = 1, 2, \ldots$ is a sequence in a metric space (X, ρ) and*

- any subsequence of t_n contains a convergent subsequence;

- $\lim_{n\to\infty} \rho(t_{n+1}, t_n) = 0$;

then the set of all limit points of t_n is connected in (X, ρ).

Now put $X=$ the set of all d.f.s defined on

$$[0,1], t_N = F_N(x) \quad \text{and} \quad \rho(t_{N+1}, t_N) = d\big(F_{N+1}(x), F_N(x)\big).$$

The limit $d\big(F_{N+1}(x), F_N(x)\big) \to 0$ follows directly from the definition $F_N(x)$, using the identity

$$\int_0^1 \big(g_1(x) - g_2(x)\big)^2 \mathrm{d}x = \int_0^1 \int_0^1 |x-y| \mathrm{d}g_1(x) \mathrm{d}g_2(y) -$$
$$- \frac{1}{2}\int_0^1 \int_0^1 |x-y| \mathrm{d}g_1(x) \mathrm{d}g_1(y) - \frac{1}{2}\int_0^1 \int_0^1 |x-y| \mathrm{d}g_2(x) \mathrm{d}g_2(y) \quad (6)$$

which holds for every d.f.s $g_1(x)$ and $g_2(x)$. Putting $g_1(x) = F_{N+1}(x)$ and $g_2(x) = F_N(x)$ we find exactly

$$\int_0^1 (F_{N+1}(x) - F_N(x))^2 \mathrm{d}x =$$
$$- \frac{1}{2(N+1)^2 N^2} \sum_{m,n=1}^{N} |x_m - x_n| + \frac{1}{(N+1)^2 N} \sum_{n+1}^{N} |x_{N+1} - x_n|, \quad (7)$$

where the right-hand side of (7) tends to zero as $N \to \infty$.

Putting $g_1(x) = F_N(x)$ and $g_2(x) = F_M(x)$ in (6) then we have

Theorem 2. *The sequence $x_n \in [0, 1)$ possesses an a.d.f. if and only if*

$$\lim_{M,N\to\infty} \left(\frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} |x_m - x_n| - \frac{1}{2M^2} \sum_{m,n=1}^{M} |x_m - x_n| \right.$$
$$\left. - \frac{1}{2N^2} \sum_{m,n=1}^{N} |x_m - x_n| \right) = 0.$$

Now, we prove (6) for every two d.f.s $g_1(x)$ and $g_2(x)$.

P r o o f. For given $g_1(x)$ and $g_2(x)$, let $x_n \in [0,1)$, $n = 1, 2, \ldots$, be a sequence such that there exist index sequences $N_1 < N_2 < \ldots$ and $M_1 < M_2 < \ldots$ such that $\lim_{k\to\infty} F_{N_k}(x) = g_1(x)$ and $\lim_{k\to\infty} F_{M_k}(x) = g_2(x)$. Such sequence x_n exists. Put $N_k = N$, $M_k = M$ and express $F_N(x) = \frac{1}{N}\sum_{n=1}^{N} c_{(x_n,1]}(x)$, $F_M(x) = \frac{1}{M}\sum_{m=1}^{M} c_{(x_m,1]}(x)$. Compute [5]

$$\begin{aligned}
&\int_0^1 \big(F_N(x) - F_M(x)\big)^2 \mathrm{d}x \\
&= \int_0^1 \left(\frac{1}{N}\sum_{n=1}^{N} c_{(x_n,1]}(x) - \frac{1}{M}\sum_{m=1}^{M} c_{(x_m,1]}(x)\right)^2 \mathrm{d}x \\
&= \frac{1}{N^2}\sum_{m,n=1}^{N}\int_0^1 c_{(x_n,1]}(x)c_{(x_m,1]}(x)\mathrm{d}x + \frac{1}{M^2}\sum_{m,n=1}^{M}\int_0^1 c_{(x_n,1]}(x)c_{(x_m,1]}(x)\mathrm{d}x \\
&- 2\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{M}\int_0^1 c_{(x_n,1]}(x)c_{(x_m,1]}(x)\mathrm{d}x. \qquad (8)
\end{aligned}$$

Since

$$\int_0^1 c_{(x_n,1]}(x)c_{(x_m,1]}(x)\mathrm{d}x = 1 - \max(x_m, x_n)$$

(8) implies

$$\begin{aligned}
&\int_0^1 \big(F_N(x) - F_M(x)\big)^2 \mathrm{d}x \qquad (9) \\
&= \int_0^1\int_0^1 \big(1 - \max(x,y)\big)\mathrm{d}F_N(x)\mathrm{d}F_N(y) \\
&+ \int_0^1\int_0^1 \big(1 - \max(x,y)\big)\mathrm{d}F_M(x)\mathrm{d}F_M(y) \\
&- 2\int_0^1\int_0^1 \big(1 - \max(x,y)\big)\mathrm{d}F_M(x)\mathrm{d}F_N(y) \\
&= \int_0^1\int_0^1 \big(1 - \max(x,y)\big)\mathrm{d}\big(F_N(x) - F_M(x)\big)\mathrm{d}\big(F_N(y) - F_M(y)\big). \qquad (10)
\end{aligned}$$

The limit of (9) by Lebesgue theorem of dominant convergence and the limit of (10) by Second Helly theorem, where

$$M = M_k, \quad N = N_k, \quad k \to \infty,$$

[5] $c_A(x)$ is the characteristic function of the set A.

gives

$$\int_0^1 \big(g_1(x)-g_2(x)\big)^2\mathrm{d}x = \int_0^1\int_0^1 \big(1-\max(x,y)\big)\mathrm{d}\big(g_1(x)-g_2(x)\big)\mathrm{d}\big(g_1(y)-g_2(y)\big). \tag{11}$$

Since

$$\max(x,y) = \frac{x+y+|x-y|}{2},$$

and

$$\int_0^1\int_0^1 1.\mathrm{d}\big(g_1(x)-g_2(x)\big)\mathrm{d}\big(g_1(y)-g_2(y)\big) = 0,$$

$$\int_0^1\int_0^1 (x+y)\mathrm{d}\big(g_1(x)-g_2(x)\big)\mathrm{d}\big(g_1(y)-g_2(y)\big) = 0,$$

we find (6) in the form

$$\int_0^1 \big(g_1(x)-g_2(x)\big)^2\mathrm{d}x = \int_0^1\int_0^1 -\frac{|x-y|}{2}\mathrm{d}\big(g_1(x)-g_2(x)\big)\mathrm{d}\big(g_1(y)-g_2(y)\big). \tag{12}$$

□

Theorem 3. *Assume that for the sequence $x_n \in [0,1)$ there exists the first moment*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{n} x_n = \alpha.$$

Then either $G(x_n)$ is singleton or $\underline{g} \notin G(x_n)$ or $\overline{g} \notin G(x_n)$.

P r o o f. We have $\int_0^1 x\mathrm{d}F_N(x) = \frac{1}{N}\sum_{n=1}^N x_n$ and by Helly Theorem, if the first moment is constant, then for every $g(x) \in G(x_n)$, we have $\int_0^1 x\mathrm{d}g(x) = \alpha$. Thus if $\underline{g}(x), \overline{g}(x) \in G(x_n)$, then $\int_0^1 x\mathrm{d}\underline{g}(x) = \int_0^1 x\mathrm{d}\overline{g}(x) = \alpha$ and from $\underline{g}(x) \le \overline{g}(x)$ for all $x \in [0,1]$ follows $\underline{g}(x) = \overline{g}(x)$ for common continuity points $x \in [0,1]$. □

Theorem 4. *Let $x_n, y_n \in [0,1)$, $n = 1,2,\ldots$. Then*

$$\frac{1}{N}\sum_{n=1}^{N}|x_n - y_n| \to 0 \Longrightarrow G(x_n) = G(y_n)$$

P r o o f. Put

$$F_N^{(1)}(x) = \frac{1}{N}\sum_{n=1}^{N} c_{[0,x)}(x_n)$$

and

$$F_N^{(2)}(x) = \frac{1}{N}\sum_{n=1}^{N} c_{[0,x)}(y_n)$$

and applying (6) we find

$$\int_0^1 \left(F_N^{(1)}(x) - F_N^{(2)}(x)\right)^2 \mathrm{d}x = \frac{1}{N^2}\sum_{m,n=1}^{N} |x_m - y_n| - \\ -\frac{1}{2}\frac{1}{N^2}\sum_{m,n=1}^{N} |x_m - x_n| - \frac{1}{2}\frac{1}{N^2}\sum_{m,n=1}^{N} |y_m - y_n|. \quad (13)$$

From

$$x_m - y_n = x_m - x_n + x_n - y_n \quad \text{and} \quad x_m - y_n = y_m - y_n + x_m - y_n$$

follows

$$|x_m - y_n| \le \frac{1}{2}|x_m - x_n| + \frac{1}{2}|y_m - y_n| + \frac{1}{2}|x_n - y_n| + \frac{1}{2}|x_m - y_m|. \quad (14)$$

Substitute (14) to (13) then we find

$$\int_0^1 \left(F_N^{(1)}(x) - F_N^{(2)}(x)\right)^2 \mathrm{d}x \le \frac{1}{N}\sum_{n=1}^{N} |y_n - x_n|. \quad (15)$$

□

Example 1. Let $\{x\}$ be the fractional part of x. For $x_n = \{\log n\}$, $n = 1, 2, \ldots$, we have the set of d.f.s

$$G(x_n) = \left\{ g_u(x) = \frac{1}{e^u}\frac{e^x - 1}{e - 1} + \frac{e^{\min(x,u)} - 1}{e^u}; u \in [0,1] \right\}, \quad (16)$$

and

$$\{\log N_k\} \to u \quad \text{implies} \quad F_{N_k}(x) \to g_u(x).$$

The lower and upper d.f. of $\log n \bmod 1$ are

$$\underline{g}(x) = \frac{e^x - 1}{e - 1}, \qquad \overline{g}(x) = \frac{1 - e^{-x}}{1 - e^{-1}},$$

and $\underline{g} \in G(x_n)$ but $\overline{g} \notin G(x_n)$. This set $G(x_n)$ was found by A. Wintner [47], also see Theorem 21.

2.2. Statistically independent sequences

G. Rauzy [33, p. 91, 4.1. Def.]:

Definition 1. Let x_n and y_n be two infinite sequences from the unit interval $[0, 1)$. The pair of sequences (x_n, y_n) is called *statistically independent* if

$$\lim_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N} f_1(x_n)f_2(y_n) - \left(\frac{1}{N}\sum_{n=1}^{N} f_1(x_n)\right)\left(\frac{1}{N}\sum_{n=1}^{N} f_2(y_n)\right)\right) = 0$$

for all continuous real functions f_1, f_2 defined on $[0, 1]$.

Theorem 5 (G. Rauzy (1976) [33]). *Two sequences* $x_n \bmod 1$ *and* $y_n \bmod 1$ *are statistically independent if and only if*

$$\lim_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i(hx_n+ky_n)} - \left(\frac{1}{N}\sum_{n=1}^{N} e^{2\pi ihx_n}\right)\left(\frac{1}{N}\sum_{n=1}^{N} e^{2\pi iky_n}\right)\right) = 0$$

for every integers $(h,k) \neq (0,0)$.

Theorem 6 (G. Rauzy [33, p. 92, 4.2. par.]). *For an arbitrary* $(x_n, y_n) \in [0,1)^2$, $n = 1, 2, \ldots$, *the sequences* x_n *and* y_n *are statistically independent if and only if*

$$\underset{g\in G(x_n,y_n)}{\forall} g(x,y) = g(x,1)g(1,y) \quad \textit{a.e. on } [0,1]^2.$$

P r o o f. For given two-dimensional sequence (x_n, y_n) put

$$F_N(x,y) = \frac{\#\{n \le N; (x_n,y_n) \in [0,x)\times[0,y)\}}{N}.$$

By Riemann-Stieltjes integration and Helly theorem, there exist a sequence of indices $N_1 < N_2 < \ldots$ and d.f. $g(x,y)$ such that

$$\frac{1}{N_k}\sum_{n=1}^{N_k} f_1(x_n)f_2(y_n) = \int_0^1\int_0^1 f_1(x)f_2(y)\mathrm{d}F_{N_k}(x,y) \to \int_0^1\int_0^1 f_1(x)f_2(x)\mathrm{d}g(x,y),$$

$$\frac{1}{N_k}\sum_{n=1}^{N_k} f_1(x_n) = \int_0^1 f_1(x)\mathrm{d}F_{N_k}(x,1) \to \int_0^1 f_1(x)\mathrm{d}g(x,1),$$

$$\frac{1}{N_k}\sum_{n=1}^{N_k} f_2(y_n) = \int_0^1 f_2(y)\mathrm{d}F_{N_k}(1,y) \to \int_0^1 f_2(y)\mathrm{d}g(1,y)$$

as $k \to \infty$. Assuming statistical independence x_n and y_n we have

$$\int_0^1\int_0^1 f_1(x)f_2(x)\mathrm{d}g(x,y) = \left(\int_0^1 f_1(x)\mathrm{d}g(x,1)\right)\left(\int_0^1 f_2(y)\mathrm{d}g(1,y)\right).$$

The integration by parts gives

$$\int_0^1\int_0^1 f_1(x)f_2(x)\mathrm{d}g(x,y) = f_1(1)f_2(1) - f_2(1)\int_0^1 g(x,1)\mathrm{d}f_1(x) - f_1(1)\int_0^1 g(1,y)\mathrm{d}f_2(y) + \int_0^1\int_0^1 g(x,y)\mathrm{d}f_1(x)\mathrm{d}f_2(y)$$

and

$\int_0^1 f_1(x)\mathrm{d}g(x,1) = f_1(1) - \int_0^1 g(x,1)\mathrm{d}f_1(x)$,

$\int_0^1 f_1(x)\mathrm{d}g(x,1) = f_2(1) - \int_0^1 g(1,y)\mathrm{d}f_2(y)$.

From it follows

$$\int_0^1 \int_0^1 g(x,y)\mathrm{d}f_1(x)\mathrm{d}f_2(y) = \left(\int_0^1 g(x,1)\mathrm{d}f_1(x)\right)\left(\int_0^1 g(1,y)\mathrm{d}f_2(y)\right)$$

for an arbitrary differentiable $f_1(x)$ and $f_2(y)$. Now, for a continuity point (x_0, y_0) of $g(x,y)$ we can select $f_1(x)$ and $f_2(y)$ such that the above implies $g(x_0,y_0) = g(x_0,1)g(1,y_0)$. □

Note that Grabner and Tichy [16] proved that the extremal discrepancy $\sup_{x,y\in[0,1]} |F_N(x,y) - F_N(x,1)F_N(1,y)|$ does not characterize statistical independence, but the L^2-discrepancy $\int_0^1\int_0^1 (F_N(x,y)-F_N(x,1)F_N(1,y))^2\mathrm{d}x\mathrm{d}y$ provides a characterization. L^2-discrepancy can be computed also by Wiener's measure $\mathrm{d}f$ (see [38]):

$$\begin{aligned}
&\int_{\mathbf{X}}\int_{\mathbf{X}} \left(\frac{1}{N}\sum_{n=1}^N f(x_n)g(y_n) - \frac{1}{N}\sum_{n=1}^N f(x_n)\frac{1}{N}\sum_{n=1}^N g(x_n)\right)^2 \mathrm{d}f\mathrm{d}g \\
&= \frac{1}{N^2}\sum_{m,n=1}^N \frac{\min(x_m,x_n)}{2}\frac{\min(y_m,y_n)}{2} + \frac{1}{N^4}\sum_{m,n,k,l=1}^N \frac{\min(x_m,x_n)}{2}\frac{\min(y_k,y_l)}{2} \\
&- \frac{2}{N^3}\sum_{m,k,l=1}^N \frac{\min(x_m,x_k)}{2}\frac{\min(y_m,y_l)}{2}. \qquad (17)
\end{aligned}$$

Theorem 7. *Let x_n and y_n be two sequences in $(0,1)^2$. If*

(i) *x_n and y_n are statistically independent;*
(ii) *x_n is u.d.;*
(iii) *all $g(x) \in G(y_n)$ are continuous;*

then the sequence $x_n + y_n \bmod 1$, $n = 1, 2, \ldots$ is u.d.

Proof. By (i) and (ii) every $g(x,y) \in G(x_n, y_n)$ has the form $g(x,y) = xg(y)$. Divide unit square $[0,1]^2$ into three parts in Fig. 1

$X_1(t) = \{(x,y) \in [0,1]; x + y < t\}$,
$X_2(t) = \{(x,y) \in [0,1]; 1 < x + y < t + 1, x \le t\}$,
$X_3(t) = \{(x,y) \in [0,1]; 1 < x + y < t + 1, x > t\}$,

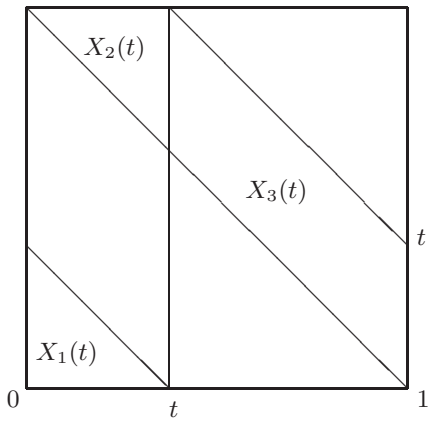


FIGURE 1. Regions $X_i(t)$

By integration

$$\int_{X_1(t)} 1.\mathrm{d}g(x,y) = \int_0^t \mathrm{d}x \int_0^{t-x} 1.\mathrm{d}g(y) = \int_0^t g(t-x)\mathrm{d}x$$

$$\int_{X_2(t)} 1.\mathrm{d}g(x,y) = \int_0^t \mathrm{d}x \int_{1-x}^1 1.\mathrm{d}g(y) = \int_0^t (1-g(1-x))\mathrm{d}x$$

$$\int_{X_3(t)} 1.\mathrm{d}g(x,y) = \int_t^1 \mathrm{d}x \int_{1-x}^{t+1-x} 1.\mathrm{d}g(y) = \int_t^1 (g(t+1-x)-g(1-x))\mathrm{d}x.$$

Thus

$$\int_{x+y \bmod 1\in[0,t)} 1\mathrm{d}g(x,y) = \int_0^t 1.\mathrm{d}x - \int_0^1 g(1-x)\mathrm{d}x + \int_0^t g(t-x)\mathrm{d}x + \int_t^1 g(t+1-x)\mathrm{d}x.$$

Now, by integrating

(j) $-\int_0^1 g(1-x)\mathrm{d}x = \int_0^1 g(x)\mathrm{d}x$,

(jj) $\int_0^t g(t-x)\mathrm{d}x = \int_0^t g(x)\mathrm{d}x$,

(jjj) $\int_t^1 g(t+1-x)\mathrm{d}x = \int_t^1 g(x)\mathrm{d}x$,

then finally we have

$$\int_{x+y \bmod 1\in[0,t)} 1\mathrm{d}g(x,y) = t. \qquad \square$$

Example 2. Let x_n and y_n be two sequences in $[0,1)$. Assume that

(i) x_n and y_n are u.d.

(ii) x_n and y_n are statistically independent.

Then by Theorem 7 the sequence $x_n + y_n \bmod 1$ is again u.d. It can be proved directly by Weyl's criterion if we prove

$$\frac{1}{N}\sum_{n=1}^{N}(\{x_n+y_n\})^k \to \frac{1}{k+1}, \quad k=1,2,\dots$$

From (ii) follows that the sequence (x_n, y_n) has a.d.f. $g(x,y) = xy$ and by Helly theorem

$$\frac{1}{N}\sum_{n=1}^{N}(\{x_n+y_n\})^k \to \int_0^1\int_0^1 (\{x+y\})^k \mathrm{d}x\mathrm{d}y, \quad k=1,2,\dots$$

Now

$$\int_0^1\int_0^1 (\{x+y\})^k \mathrm{d}x\mathrm{d}y = \iint_{0\le x+y\le 1}(x+y)^k\mathrm{d}x\mathrm{d}y + \iint_{1\le x+y\le 2}(x+y-1)^k\mathrm{d}x\mathrm{d}y$$

which is $\frac{1}{k+1}$ and the proof is finished.

Theorem 8 (G. Rauzy [33]). *Let $x_n \in [0,1)$, $n = 1,2,\dots$, be u.d. sequence. Then x_n and $\log n \bmod 1$ are statistically independent, i.e., every $g(x,y) \in G(x_n, \{\log n\})$ has the form $g(x,y) = x.g(1,y)$.*

Proof. Let $N \in [e^K, e^{K+1})$ i.e., $N = e^{K+\theta_N}$, and divide $n \le N$ to the subsets $n \in [e^k, e^{k+1})$, $k \le K$. For such n we have $\{\log n\} \in [0,y) \iff n \in [e^k, e^{k+y})$. For $n \in [e^k, e^{k+y})$ we ask the number of $x_n \in [0,x)$ which is $x(e^{k+y} - e^k) + O(e^k D_{e^k} + e^{k+y} D_{e^{k+y}})$. Omitting integer parts here we use discrepancy D_M of the initial string $x_1, x_2, \dots, x_M$ (for definition of discsrepancy, see [44, 1–40]) and the formula $A([0,x); M; x_n) = xM + O(MD_M)$. [6] Thus

$$\frac{A\big([0,x)\times[0,y); N; (x_n, \{\log n\})\big)}{N}$$
$$= \frac{\sum_{k=0}^{K-1} x(e^{k+y} - e^k) + x(e^{K+\min(y,\theta_N)} - e^K) + O(\sum_{k=0}^{K} e^k D_{e^k} + e^{k+y} D_{e^{k+y}})}{N}.$$

[6] O-constant can be put $= 1$ and in the interval $[e^K, e^{K+\min(y,\theta_N)}]$, for simplification, an error term we put $O(e^k D_{e^K} + e^{K+y} D_{e^{K+y}})$.

As $N \to \infty$ and $\theta_N \to u$, we have

$$\frac{\sum_{k=0}^{K-1} x(e^{k+y}-e^k)}{N} = \frac{\sum_{k=0}^{K-1} x(e^{k+y}-e^k)}{\sum_{k=0}^{K-1}(e^{k+y}-e^k)} \frac{\sum_{k=0}^{K-1}(e^{k+y}-e^k)}{N} \to x\frac{e^y-1}{e-1}\frac{1}{e^u},$$

$$\frac{x(e^{K+\min(y,\theta_N)}-e^K)}{N} \to x\frac{e^{\min(y,u)}-1}{e^u},$$

$$\frac{O(\sum_{k=0}^{K} e^k D_{e^k} + e^{k+y} D_{e^{k+y}})}{N} = O\left(\frac{\sum_{k=0}^{K} e^k D_{e^k} + e^{k+y} D_{e^{k+y}}}{\sum_{k=0}^{K}(e^{k+y}-e^k)}\right) \to 0.$$

In the final parenthesis we have used $\frac{e^k D_{e^k} + e^{k+y} D_{e^{k+y}}}{e^{k+y}-e^k} \to 0$ as $k \to \infty$. Collected all above results we have

$$\frac{A([0,x)\times[0,y);N;(x_n,\{\log n\}))}{N} \to x\left(\frac{e^y-1}{e-1}\frac{1}{e^u} + \frac{e^{\min(y,u)}-1}{e^u}\right) = xg_u(y),$$

where $g_u(y)$ is the same as in (16). □

In 2011 Y. Ohkubo [27] proved that the function $\log n$ can be replaced by $\log(n\log n)$ in Theorem 8.

Theorem 9. *An arbitrary u.d. sequence* $x_n \bmod 1$ *and* $\log(n\log n) \bmod 1$ *are statistically independent.*

Proof. By G. Rauzy Theorem 5 two sequences $x_n \bmod 1$ and $y_n \bmod 1$ are statistically independent if and only if

$$\lim_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i(hx_n+ky_n)} - \left(\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i h x_n}\right)\left(\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i k y_n}\right)\right) = 0$$

for every integers h and k. Now, by Abel partial summation we obtain

$$\begin{aligned}
&\sum_{n=1}^{N} e^{2\pi i(hx_n + k\log(n\log n))} \\
&= \sum_{n=1}^{N-1}\left(e^{2\pi i k\log(n\log n)} - e^{2\pi i k\log((n+1)\log(n+1))}\right)\sum_{j=1}^{n} e^{2\pi i h x_n} \\
&\quad + e^{2\pi i k\log(N\log N)}\sum_{j=1}^{N} e^{2\pi i h x_n}
\end{aligned}$$

and

$$\left|e^{2\pi i k\log(n\log n)} - e^{2\pi i k\log((n+1)\log(n+1))}\right| \le 2\pi|k|\frac{(\log n)+1}{n\log n}.$$

Thus

$$\left|\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i(hx_n+k\log(n\log n))}\right| \leq \frac{1}{N}\sum_{n=1}^{N-1} 2\pi|k|\frac{(\log n)+1}{n\log n}n\left|\frac{1}{n}\sum_{j=1}^{n} e^{2\pi ihx_j}\right| + \left|\frac{1}{N}\sum_{j=1}^{N} e^{2\pi ihx_j}\right|$$

which tends to 0. □

Using Theorem 9 Ohkubo [27] proved that in Theorem 8 the $\log n$ can be instead by $\log p_n$, where p_n are sequence of all primes.

Theorem 10. *Let $x_n \in [0,1)$, $n = 1,2,\ldots$, be u.d. sequence. Then x_n and $\log p_n \bmod 1$ are statistically independent.*

P r o o f. Firstly he proved that

$$\log p_n = \log(n\log n) + o\left(\frac{\log\log n}{\log n}\right) + O\left(\frac{1}{\log p_n}\right). \tag{18}$$

Then Ohkubo used

Let (x_n, y_n) and (x'_n, y'_n), $n = 1,2,\ldots$ be two-dimensional sequences. Assume:

(i) $|x_n - x'_n| \to 0$ and $|y_n - y'_n| \to 0$.

(ii) Every d.f. $g(x,y) \in G\big((x_n,y_n)\big)$ is continuous in $(0,0)$,$(0,1)$, $(1,0)$ and $(1,1)$.

Then $G\big((x_n,y_n)\big) = G\big((x'_n,y'_n)\big)$.

Then the limit

$$\lim_{n\to\infty}\big(\log p_n - \log(n\log n)\big) = 0,$$

given by (18), implies

$$G\big((x_n,\{\log p_n\})\big) = G\big((x_n,\{\log(n\log n)\})\big) = G\big((x_n,\{\log n\})\big). \tag{19}$$

Proof of (18). He starts with the prime number theorem of the form

$$\pi(x) = \frac{x}{\log x\ -1} + O\left(\frac{x}{(\log x)^3}\right). \tag{20}$$

This implies

$$\frac{p_n}{n} = \log p_n\ -1 + O\left(\frac{1}{\log p_n}\right).$$

Then he used (see [43])

$$\frac{p_n}{n} = \log n + (\log\log n\ -1) + o\left(\frac{\log\log n}{\log n}\right)$$

which implies (18). □

Some generalization:

Example 3. J. Coquet and P. Liardet [7]: Given an integer $q \geq 2$, a real number θ and a real polynomial $p(x)$, let

(i) $x_n = \theta q^n \bmod 1$,

(ii) $y_n = p(n) \bmod 1$,

(iii) $\mathbf{x}_n = (x_{n+1}, \ldots, x_{n+s})$ and $\mathbf{y}_n = (y_{n+1}, \ldots, y_{n+s})$.

If x_n is u.d. (i.e., θ is normal in the base q), then for every $s = 1, 2, \ldots$, the sequence

$$(\mathbf{x}_n, \mathbf{y}_n), \quad n = 1, 2, \ldots,$$

has d.f.s $g(\mathbf{x}, \mathbf{y}) \in G\big((\mathbf{x}_n, \mathbf{y}_n)\big)$ of the form $g(\mathbf{x}, \mathbf{y}) = g_1(\mathbf{x})g_2(\mathbf{y})$ for some $g_1(\mathbf{x}) \in G(\mathbf{x}_n)$ and $g_2(\mathbf{y}) \in G(\mathbf{y}_n)$, i.e., the sequences x_n and y_n are *completely statistically independent*.

2.3. Statistical limit

H. Fast [10] and I.J. Schoenberg [34] defined, independently:

Definition 2. The sequence x_n is said to be *statistically convergent* to the number α provided that for each $\varepsilon > 0$,

$$\lim_{N\to\infty} \frac{1}{N} \#\{n \leq N; |x_n - \alpha| \geq \varepsilon\} = 0.$$

• Fast [10] mentioned: A sequence x_n is statistically convergent to α if and only if there exists a sequence of indices k_n of the asymptotic density $d(k_n) = 1$ such that $\lim_{n\to\infty} x_{k_n} = \alpha$ in the standard sense.

• Let us consider *one-jump function* $c_\alpha(x)$ which has a jump of size 1 for α.

Theorem 11 (I.J. Schoenberg [34])**.** *The sequence $x_n \in [0, 1)$ is statistically convergent to the number $\alpha \in [0, 1]$ if and only if the sequence x_n admits the asymptotic distribution function $c_\alpha(x)$.*

Example 4. [44, p. 2–192, 2.20.18]: Let $\operatorname{ord}_p(n) = \alpha$ for $p^\alpha \| n$. If p stands for a prime, then the sequence

$$\log p \frac{\operatorname{ord}_p(n)}{\log n}, \quad n = 2, 3, \ldots,$$

is dense in $[0, 1]$ and has the a.d.f. $c_0(x)$, and thus statistically converge to zero.

Theorem 12 ([36]). *The sequence $x_n \in [0,1)$ possesses a statistical limit if and only if*

$$\lim_{M,N\to\infty} \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} |x_m - x_n| = 0.$$

P r o o f. Let $F_{M_{i_k}}(x) \to g_1(x)$ and $F_{N_{j_k}}(x) \to g_2(x)$. Applying the Helly-Bray theorem we find

$$\lim_{k\to\infty} \int_0^1 \int_0^1 |x-y| \mathrm{d}F_{M_{i_k}}(x) \mathrm{d}F_{N_{j_k}}(y) = \int_0^1 \int_0^1 |x-y| \mathrm{d}g_1(x) \mathrm{d}g_2(y).$$

By Riemann-Stieltjes integration we obtain

$$\int_0^1 \int_0^1 |x-y| \mathrm{d}F_{M_{i_k}}(x) \mathrm{d}F_{N_{j_k}}(y) = \frac{1}{M_{i_k} N_{j_k}} \sum_{m=1}^{M_{i_k}} \sum_{n=1}^{N_{j_k}} |x_m - x_n|.$$

Thus

$$\int_0^1 \int_0^1 |x-y| \mathrm{d}g_1(x) \mathrm{d}g_2(y) = 0$$

which gives $g_1(x) = g_2(x) = c_\alpha(x)$ a.e. for some $\alpha \in [0,1]$. □

2.4. Statistical limit points

Following the concept od statistical convergence J. A. Fridy [14] introduced:

Definition 3. A real number x is said to be *a statistical limit point* of the sequence x_n if there exists a subsequence x_{k_n}, $n = 1,2,\ldots$, such that $\lim_{n\to\infty} x_{k_n} = x$ and the set of indices k_n has a positive upper asymptotic density.

Fridy studied the set $\Lambda(x_n)$ of all such points. Inspired by I.J. Schoenberg [34], P. Kostyrko, M. Mačaj, T. Šalát and O. Strauch [21] was found:

Theorem 13. *The set $\Lambda(x_n)$, for $x_n \in [0,1)$ $n = 1,2,\ldots$, coincides with the set of all discontinuity points of d.f.s $g(x) \in G(x_n)$.*

From it follows:

(i) Let x_n be a sequence of real numbers. If for every $k = 1,2,\ldots$ the difference sequence $x_{n+k} - x_n$, $n = 1,2,\ldots$ has $\Lambda(x_{n+k} - x_n) = \emptyset$, then $\Lambda(x_n) = \emptyset$.

(ii) For u.d. sequence x_n we have $\Lambda(x_n \bmod 1) = \emptyset$.

(iii) $\omega_h = \limsup_{N\to\infty} \left| \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i h x_n} \right|^2$ for $h = 1,2,\ldots$ and assume that $\lim_{H\to\infty} \frac{1}{H} \sum_{h=1}^{H} \omega_h = 0$. The every $g(x) \in G(x_n)$ is a continuous, thus $\Lambda(x_n) = \emptyset$.

Example 5. By Example 1 every d.f. $g(x) \in G(\log n \bmod 1)$ is continuous, thus we have $\Lambda(\log n \bmod 1) = \emptyset$. More generally, for $x_n = c \log n \bmod 1$, $c \neq 0$, we have

$$\omega_h = \frac{1}{4\pi^2 h^2 c^2 + 1}$$

which implies $\lim_{H\to\infty} \frac{1}{H} \sum_{h=1}^{H} \omega_h = 0$ and thus by (iii) $\Lambda(c \log n \bmod 1) = \emptyset$, again.

Example 6. By [36] starting with $\log\log n \bmod 1$ all the sequences of iterate logarithm $\log\log \ldots \log n \bmod 1$ have

$$G(\log\log \ldots \log n \bmod 1) = \{c_\alpha(x); \alpha \in [0,1]\} \cup \{h_\alpha(x); \alpha \in [0,1]\}.$$

Here $c_\alpha(x)$ is one-step d.f. for which

$$c_\alpha(x) = 0 \quad \text{for} \quad x \in [0,\alpha], \; c_\alpha(x) = 1 \quad \text{for} \quad x \in (\alpha,1] \quad \text{and} \quad h_\alpha : [0,1] \to [0,1]$$

is a constant distribution function, where $h_\alpha(0) = 0$, $h_\alpha(1) = 1$, and $h_\alpha(x) = \alpha$ if $x \in (0,1)$. Thus we have $\Lambda(\log\log \ldots \log n \bmod 1) = [0,1]$.

Example 7. Let $\alpha = \frac{p}{q}\pi$, where p and q are positive integers and g.c.d. $(p,q) = 1$. By D. Berend, M. D. Boshernitzan, and G. Kolesnik [6] the sequence

$$x_n = n \cos(n \cos n\alpha) \bmod 1, \quad n = 1, 2, \ldots$$

has $G(x_n) = \{g(x)\}$, where

$$g(x) = \begin{cases} x & \text{if } q \text{ is odd}, \\ \left(1 - \frac{1}{q}\right) x + \frac{1}{q} c_0(x) & \text{if } q \text{ is even}, \end{cases}$$

and $c_0(x)$ is the one-jump d.f. which the jump in 0. This implies

$$\Lambda(x_n) = \begin{cases} \emptyset & \text{if } q \text{ is odd}, \\ \{0\} & \text{if } q \text{ is even}. \end{cases}$$

2.5. Uniformly maldistributed sequences

G. Myerson [24]:

Definition 4. The sequence $x_n \in [0,1)$ $n = 1, 2, \ldots$, is said to be *uniformly maldistributed* (u.m.) if for every nonempty proper subinterval $I \subset [0,1]$ we have both

$$\liminf_{N\to\infty} \frac{1}{N} \#\{n \leq N; x_n \in I\} = 0 \text{ and } \limsup_{N\to\infty} \frac{1}{N} \#\{n \leq N; x_n \in I\} = 1.$$

He mentioned that the first condition is superfluous, and showed that

Example 8. The sequence $x_n = \{\log\log n\}$ of fractional parts of the iterated logarithm is u.m.

In [36] is proved: Let $c_\alpha(x)$ is one-step d.f. for which $c_\alpha(x) = 0$ for $x \in [0, \alpha]$ and $c_\alpha(x) = 1$ for $x \in (\alpha, 1]$.

Theorem 14. *The sequence x_n is u.m. if and only if*

$$\{c_\alpha(x); \alpha \in [0,1]\} \subset G(x_n).$$

Example 9. By Example 6, starting with $x_n = \{\log\log n\}$, all the sequences $x_n = \{\log\log\ldots\log n\}$, $n = n_0, n_0+1, \ldots$ are u.m.

Thus, in the theory of uniform maldistribution we need not consider d.f.s other than one-jump d.f. $c_\alpha(x)$ which has a jump of size 1 at α. This suggests the definition (see [36]):

Definition 5. The sequence x_n is said to be *uniformly maldistributed in the strict sense* (u.m.s.) if $G(x_n) = \{c_\alpha(x); \alpha \in [0,1]\}$.

Theorem 15. *For every sequence $x_n \in [0,1)$ we have*

$$G(x_n) \subset \{c_\alpha(x); \alpha \in [0,1]\} \iff \lim_{N\to\infty} \frac{1}{N^2} \sum_{m,n=1}^{N} |x_m - x_n| = 0.$$

Moreover, if $G(x_n) \subset \{c_\alpha(x); \alpha \in [0,1]\}$, then $G(x_n) = \{c_\alpha(x); \alpha \in I\}$, where I is a closed subinterval of $[0,1]$ which can be found as

$$I = \left[\liminf_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} x_n,\ \limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} x_n\right],$$

and the length $|I|$ of I can also be found as

$$|I| = \limsup_{M,N\to\infty} \frac{1}{MN} \sum_{m=1}^{M}\sum_{n=1}^{N} |x_m - x_n|.$$

The following theorem is immediately evident from the preceding,

Theorem 16. *The sequence $x_n \in [0,1)$ is u.m.s. if and only if*

$$\lim_{N\to\infty} \frac{1}{N^2} \sum_{m,n=1}^{N} |x_m - x_n| = 0 \text{ and } \limsup_{M,N\to\infty} \frac{1}{MN} \sum_{m=1}^{M}\sum_{n=1}^{N} |x_m - x_n| = 1,$$

or alternatively $\limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} x_n - \liminf_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} x_n = 1$.

Example 10. Let x_n, $n = 1, 2, \ldots$ be defined as

$$x_n = \left\{ 1 + (-1)^{\left[\sqrt{[\sqrt{\log_2 n}]}\right]} \left\{ \sqrt{[\sqrt{\log_2 n}]} \right\} \right\},$$

were $[x]$ denotes the integral part and $\{x\}$ the fractional part of x. Then

$$G(x_n) = \{c_\alpha(x); \alpha \in [0, 1]\}.$$

2.6. Benford's law

The first digit problem:

2.6.1. Historical notes

An infinite sequence $x_n \geq 1$, $n = 1, 2, \ldots$, of real numbers satisfies *Benford's law*, if the frequency (the asymptotic density) of occurrences of a given first digit a, when x_n is expressed in the decimal form is given by $\log_{10}\left(1 + \frac{1}{a}\right)$ for every $a = 1, 2, \ldots, 9$ (0 as a possible first digit is not admitted).

It was S. Newcomb (1881), who firstly noted "*That the ten digits do not occur with equal frequency must be evident to anyone making use of logarithm tables*". F. Benford (1938) compared the empirical frequency of occurrences of a with $\log_{10}((a+1)/a)$ in twenty different tables. Since x_n has the first digit a if and only if

$$\log_{10} x_n \bmod 1 \in [\log_{10} a, \log_{10}(a+1)),$$

Benford's law for x_n follows from the uniform distribution of $\log_{10} x_n \bmod 1$. For the asymptotic density of the second-place digit b he found

$$\sum_{a=1}^{9} \log_{10}\left(1 + \frac{1}{10a+b}\right).$$

F. Benford rediscovered Newcomb's observation from (1881). P. Diaconis [8] suggested the following generalization:

2.6.2. Generalization of Benford's law

Let $b \geq 2$ be an integer considered as a base for the development of a real number $x > 0$ and $M_b(x)$ be the mantissa of x defined by $x = M_b(x) \times b^{n(x)}$ such that $1 \leq M_b(x) < b$ holds, where $n(x)$ is a uniquely determined integer. Let $K = k_1 k_2 \cdots k_r$ be a positive integer expressed in the base b , that is

$$K = k_1 \times b^{r-1} + k_2 \times b^{r-2} + \cdots + k_{r-1} \times b + k_r,$$

where $k_1 \neq 0$ and at the same time $K = k_1 k_2 \cdots k_r$ is considered as an $r-$consecutive block of digits in the base b. Note that for x of the type $x = 0.00 \cdots$

$\cdots 0k_1k_2\cdots k_r\cdots$, $k_1>0$, we have $M_b(x)=k_1.k_2\cdots k_r\cdots$ and the first zero digits are omitted. Thus arbitrary $x>0$ has the first r-digits, starting a non-zero digit, equal to $k_1k_2\cdots k_r$ if and only if [7]

$$k_1.k_2\cdots k_r\le M_b(x)<k_1.k_2\cdots(k_r+1). \tag{21}$$

Since $\log_b M_b(x)=\log_b x \bmod 1$ the inequality (21) is equivalent to

$$\log_b\left(\frac{K}{b^{r-1}}\right)\le \log_b x \bmod 1<\log_b\left(\frac{K+1}{b^{r-1}}\right). \tag{22}$$

Here we use the shorthand notation $\frac{K}{b^{r-1}}=k_1.k_2\cdots k_r$.

Definition 6. A sequence x_n, $n=1,2,\dots$, of positive real numbers satisfies *Benford's law* (abbreviated to B.L.) [8] in base b, if for every $r=1,2,\dots$ and every r-digits integer $K=k_1k_2\cdots k_r$ we have the density

$$\lim_{N\to\infty}\frac{\#\{n\le N;\ \text{leading block of } r \text{ digits (beginning with } \ne 0) \text{ of } x_n=K\}}{N}$$
$$=\log_b\left(\frac{K+1}{b^{r-1}}\right)-\log_b\left(\frac{K}{b^{r-1}}\right). \tag{23}$$

From (22) and from definition (23) it follows immediately:

Theorem 17. *A sequence x_n, $n=1,2,\dots$, of positive real numbers satisfies B.L. in base b if and only if the sequence* $\log_b x_n \bmod 1$ *is u.d. in* $[0,1)$.

Example 11. The sequence of Fibonacci numbers F_n, factorials $n!$, and n^n, and n^{n^2} satisfy Benford's law, but the sequence n, and all primes p_n does not, see [44, 2.12.26], [3].

P. Diaconis [8] and A. I. Pavlov [29] have been the first, who applied uniform distribution theory to B.L. For instance:

(i) By the criterion of P.B Kennedy [44, p. 2–13, 2.2.9], P. Diaconis proved: *If a sequence $x_n>0$, $n=1,2,\dots$, satisfies B.L. in the base b, then*

$$\limsup_{n\to\infty} n\left|\log\frac{x_{n+1}}{x_n}\right|=\infty. \tag{24}$$

(ii) Applying van der Corput difference theorem [22, p. 26, Th.3.1] A.I. Pavlov proved: *Assume $x_n>0$, $n=1,2,\dots$. If for every $k=1,2,\dots$ the ratio sequence $\frac{x_{n+k}}{x_n}$, $n=1,2,\dots$, satisfies B.L. in the base b, then the original sequence x_n, $n=1,2,\dots$ also satisfies B.L. in the base b.*

[7] If $k_1=k_2=\cdots=k_r=b-1$ then we have $k_1.k_2\cdots(k_r+1)=b$.

[8] precisely known as generalized or strong

(iii) In [3] we have: *The positive sequences* x_n *and* $\frac{1}{x_n}$, $n = 1, 2, \ldots$ *satisfy B.L. in the base* b *simultaneously.* Proof: Both two sequences u_n and $-u_n$ are u.d. mod1 simultaneously, since their Weyl's sums are complex conjugate each other.

(iv) *The positive sequences* x_n *and* nx_n, $n = 1, 2, \ldots$ *satisfy B.L. in the base* b *simultaneously.* Proof: Both two sequences u_n and $u_n + \log n$ are u.d. mod1 simultaneously, see [44, p. 2–27, 2.3.6.].

(v) *Assume that a sequence* $0 < x_1 \leq x_2 \leq \ldots$ *satisfies B.L. in the integer base* $b > 1$. *Then*

$$\lim_{n\to\infty} \frac{\log x_n}{\log n} = \infty. \tag{25}$$

Proof: It follows from the theorem of H. Niederreiter [44, p. 2–12, 2.2.8] that every monotone u.d. sequence $u_n \bmod 1$ must satisfy $\lim_{n\to\infty} \frac{|u_n|}{\log n} = \infty$. Here, it suffices to put $u_n = \log x_n$, instead of $u_n = \log_b x_n$.

(vi) *For a sequence* $x_n > 0$, $n = 1, 2, \ldots$, *assume that*
(i) $\lim_{n\to\infty} x_n = \infty$ *monotonically,*
(ii) $\lim_{n\to\infty} \log \frac{x_{n+1}}{x_n} = 0$ *monotonically.*
Then the sequence x_n *satisfies B.L. in every base* b *if and only if*

$$\lim_{n\to\infty} n \log \frac{x_{n+1}}{x_n} = \infty. \tag{26}$$

Proof: It follows from Fejér's difference theorem in the form in [44, p. 2–13, 2.2.11].

(vii) [44, p. 2–14, 2.2.12] implies: *Let* $x_n > 0$ *be a sequence, which satisfies* $\lim_{n\to\infty} \log_b \frac{x_{n+1}}{x_n} = \theta$ *with* θ *is irrational. Then* x_n *satisfies B.L. in the base* b.

V. Baláž, K. Nagasaka and O. Strauch [3] study d.f.s of a sequence $x_n \in (0, 1)$ which satisfy B.L. Using the Fig. 1

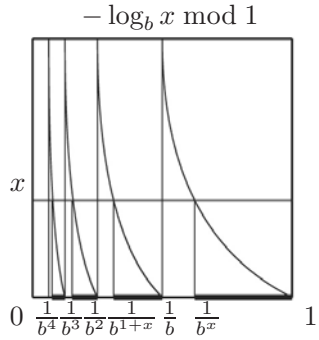


Figure 1: Intervals $f_i^{-1}([0, x))$, $i = 0, 1, 2, \ldots$

and that $f_i^{-1}([0,x)) = \left(\frac{1}{b^{i+x}}, \frac{1}{b^i}\right]$ they proved:

Theorem 18. *Let x_n, $n = 1, 2, \ldots$, be a sequence in $(0,1)$ and $G(x_n)$ be the set of all d.f.s of x_n. Assume that every d.f. $g(x) \in G(x_n)$ is continuous at $x = 0$. Then the sequence x_n satisfies B.L. in the base b if and only if for every $g(x) \in G(x_n)$ we have*

$$x = \sum_{i=0}^{\infty} \left(g\left(\frac{1}{b^i}\right) - g\left(\frac{1}{b^{i+x}}\right) \right) \text{ for } x \in [0,1]. \tag{27}$$

Example 12. We present the following solutions of (27):

$$g(x) = \begin{cases} x & \text{if } x \in \left[0, \frac{1}{b}\right], \\ 1 + \frac{\log x}{\log b} + (1-x)\frac{1}{b-1} & \text{if } x \in \left[\frac{1}{b}, 1\right]. \end{cases}$$

$$g^*(x) = \begin{cases} 0 & \text{if } x \in \left[0, \frac{1}{b^2}\right], \\ 2 + \frac{\log x}{\log b} & \text{if } x \in \left[\frac{1}{b^2}, \frac{1}{b}\right], \\ 1 & \text{if } x \in \left[\frac{1}{b}, 1\right] \end{cases}$$

$$g^{**}(x) = \begin{cases} 0 & \text{if } x \in \left[0, \frac{1}{b^3}\right], \\ 3 + \frac{\log x}{\log b} & \text{if } x \in \left[\frac{1}{b^3}, \frac{1}{b^2}\right], \\ 1 & \text{if } x \in \left[\frac{1}{b^2}, 1\right] \end{cases}$$

If H is the set of all $t_1 g(x) + t_2 g^*(x) + t_3 g^{**}(x)$, $t_1 + t_2 + t_3 = 1$, $t_1, t_2, t_3 \geq 0$, then there exists a sequence x_n such that $G(x_n) = H$ (see 2.1), and this x_n satisfies B.L.

By the following Example 13, there exist an integer sequences $x_n \in \mathbb{N}$ which satisfies B.L. for an arbitrary base b. By the following Theorem 19 no such real sequence $x_n \in [0,1)$ exists.

Example 13. By [44, p. 2–117, 2.12.14], the sequence

$$\alpha n \log^\tau n \bmod 1, \quad \alpha \neq 0,\ 0 < \tau \leq 1,$$

is u.d. From this follows that $x_n = n^n$ satisfies B.L. for an arbitrary integer base b, because $\log_b n^n = n \log n \frac{1}{\log b}$.

In [3] there is proved:

Theorem 19. *For a sequence $x_n \in (0,1)$, $n = 1, 2, \ldots$, assume that every d.f. $g(x) \in G(x_n)$ is continuous at $x = 0$. Then there exist only finitely many different integer bases b for which the sequence x_n satisfies B.L. simultaneously. Moreover, if the sequence x_n satisfies B.L. in base b, and for some $k = 1, 2, \ldots$ there exists kth integer root $\sqrt[k]{b}$, then x_n satisfies B.L. also in base $\sqrt[k]{b}$.*

As it is well known that the increasing sequence of all positive integers $1, 2, 3, \ldots$ does not satisfy B.L. in every base $b \geq 2$. It follows from the fact that $\log_b n \bmod 1$ is not u.d. For a density of n for which r initial digits are $K = k_1 k_2 \ldots k_r$, A.I. Pavlov [29] proved that

$$\liminf_{N\to\infty} \frac{\#\{n \leq N; n \text{ has the first } r \text{ digits } = K\}}{N} = \frac{1}{K(b-1)}, \tag{28}$$

$$\limsup_{N\to\infty} \frac{\#\{n \leq N; n \text{ has the first } r \text{ digits } = K\}}{N} = \frac{b}{(K+1)(b-1)}. \tag{29}$$

Using the theory of d.f.s boundaries (28) and (29)are extended in [3] to the following (30): By G. Pólya and G. Szegö [32] the d.f.s of $\log_b n \mod 1$ is of the form (see also Theorem 21)

$$g_u(x) = \frac{b^{\min(x,u)} - 1}{b^u} + \frac{1}{b^u}\frac{b^x - 1}{b-1},$$

where the parameter u runs $[0, 1]$ and by [15], for increasing sequence N_i, $i = 1, 2, \ldots$, we have

$$\log_b N_i \bmod 1 \to u \Longrightarrow F_{N_i}(x) \to g_u(x).$$

Thus

$$\frac{\#\{n \leq N_i; n \text{ has the first } r \text{ digits } = K\}}{N_i} \to \big(g_u(x_2) - g_u(x_1)\big) \tag{30}$$

as $i \to \infty$, $x_1 = \log_b(k_1.k_2k_3\ldots k_r)$, $x_2 = \log_b\big(k_1.k_2k_3\ldots(k_r+1)\big)$, and the minimum is appeared in $u = x_1$ and the maximum in $u = x_2$.

2.6.3. General scheme of solution of the First Digit Problem

Many authors think that if the sequence x_n does not satisfy B.L., then the relative density of indices n for which the b-expansion of x_n start with leading digits $K = k_1 k_2 \ldots k_r$

$$\frac{1}{N}\#\left\{n \leq N; \log_b\left(\frac{K}{b^{r-1}}\right) \leq \{\log_b x_n\} < \log_b\left(\frac{K+1}{b^{r-1}}\right)\right\}.$$

do not follow any distribution in the sense of natural density, see S. Eliahou, B. Massé and D. Schneider (2013). These authors as an alternate result shown that the sequence $\log_{10} n^r \bmod 1$, $n = 1, 2, \ldots$, and the sequence $\log_{10} p_n^r \bmod 1$ $n = 1, 2, \ldots$, p_n are all prime numbers, have the discrepancy $O(r^{-1})$. Thus, for $r \to \infty$, these sequences tends to uniform distribution and thus n^r and p_n^r tends to B.L. Using theory of d.f.'s we can give the following general solution of the first digit problem (see [28]):

Theorem 20. *Let* $g(x) \in G(\log_b x_n \bmod 1)$ *and* $\lim_{i\to\infty} F_{N_i}(x) = g(x)$. *Then*

$$\lim_{N_i\to\infty} \frac{\#\{n \le N_i;\ \textit{first } r \textit{ digits (starting a non-zero digit) of } x_n = K\}}{N_i}$$
$$= g\left(\log_b\left(\frac{K+1}{b^{r-1}}\right)\right) - g\left(\log_b\left(\frac{K}{b^{r-1}}\right)\right).$$

2.6.4. Distribution functions of sequences involving logarithm

Distribution functions of $\log_b x_n \bmod 1$ we need in generalized B.L., Theorem 20. It can be computed by following theorem: [9]

Assumptions. Let the real-valued function $f(x)$ be strictly increasing for $x \ge 1$ and let

$f^{-1}(x)$ be its inverse function and

$F_N(x) = \frac{\#\{n\le N; f(n) \bmod 1 \in [0,x)\}}{N}$ for $x \in [0,1]$.

Assume that

(i) $\lim_{x\to\infty} f'(x) = 0$,

(ii) $\lim_{k\to\infty} f^{-1}(k+1) - f^{-1}(k) = \infty$,

(iii) $\lim_{k\to\infty} \frac{f^{-1}(k+w(k))}{f^{-1}(k)} = \psi(w)$ for every sequence $w(k) \in [0,1]$ for which $\lim_{k\to\infty} w(k) = w$, where this limit defines the function
$$\psi : [0,1] \to [1, \psi(1)],$$

(iv) $\psi(1) > 1$.

Theorem 21 ([43]). *Then* (i)–(iv) *imply*

$$G(f(n) \bmod 1) = \Big\{g_w(x) = \frac{1}{\psi(w)}\frac{\psi(x)-1}{\psi(1)-1} + \frac{\min(\psi(x),\psi(w))-1}{\psi(w)}; w \in [0,1]\Big\}.$$

Now, if $w(i) = \{f(N_i)\} \to w$, *then* $F_{N_i}(x) \to g_w(x)$ *for every* $x \in [0,1]$.

Theorem 22 (Y. Ohkubo [27]). *Then* (i)–(iv) *imply*

$$G(f(p_n) \bmod 1) = \left\{g_w(x) = \frac{1}{\psi(w)}\frac{\psi(x)-1}{\psi(1)-1} + \frac{\min(\psi(x),\psi(w))-1}{\psi(w)}; w \in [0,1]\right\}.$$

Now, if $w(i) = \{f(p_{N_i})\} \to w$, *then* $F_{N_i}(x) \to g_w(x)$ *for every* $x \in [0,1]$.

Example 14 (Example of natural numbers). For the sequence

$f(n) = \log_b n^s$, $n = 1, 2, \ldots$, $f^{-1}(x) = b^{\frac{x}{s}}$,

$\lim_{k\to\infty} \frac{f^{-1}(k+w)}{f^{-1}(k)} = \frac{b^{\frac{k+w}{s}}}{b^{\frac{k}{s}}} = b^{\frac{w}{s}} = \psi(w)$, and by Theorem 21

[9] Theorem 21 is a simple version of Theorems 34, 35 and 36.

$G(\log_b n^s \bmod 1) = \{g_w(x) = \frac{1}{b^{\frac{w}{s}}} \frac{b^{\frac{x}{s}}-1}{b^{\frac{1}{s}}-1} + \frac{\min(b^{\frac{x}{s}}, b^{\frac{w}{s}})-1}{b^{\frac{w}{s}}}; w \in [0,1]\}$.

If $\lim_{i\to\infty}\{f(N_i)\} = \lim_{i\to\infty}\{\log_b(N_i^s)\} = w$, then

$$\begin{aligned}
&\lim_{i\to\infty} \frac{\#\{n \le N_i;\ \text{first } r \text{ digits of } n^s \text{ are } k_1k_2\dots k_r\}}{N_i}\\
&= g_w\big(\log_b k_1.k_2k_3\dots(k_r+1)\big) - g_w\big(\log_b k_1.k_2k_3\dots k_r\big)\\
&= \frac{b^{(\log_b k_1.k_2\dots(k_r+1))/s}-1}{b^{1/s}-1} - \frac{b^{(\log_b k_1.k_2\dots k_r)/s}-1}{b^{1/s}-1}\\
&= \frac{(k_1.k_2k_3\dots(k_r+1))^(1/s) - (k_1.k_2k_3\dots k_r))^(1/s)}{b^{1/s}-1},
\end{aligned}$$

where we assume $N_i = b^i$ which gives $\lim_{i\to\infty}\{\log_b(b^{is}\} = 0 = w$.

Example 15 (Example of primes). $f(p_n) = \log_b p_n^s$, $n = 1, 2, \dots,$ p_n is the nth prime.

$$f(x) = \log_b x^s,$$

$$G(\log_b p_n^s \bmod 1) = \{g_w(x) = \frac{1}{b^{\frac{w}{s}}} \frac{b^{\frac{x}{s}}-1}{b^{\frac{1}{s}}-1} + \frac{\min(b^{\frac{x}{s}}, b^{\frac{w}{s}})-1}{b^{\frac{w}{s}}}; w \in [0,1]\}.$$

If $N_i = \pi(b^{\frac{i}{s}}) + 1$, then $\lim_{\imath\to\infty}\{f(p_{N_i})\} = 0$ (cf. [28]) and $g_0(x) = \frac{b^{x/s}-1}{b^{1/s}-1}$. Thus

$$\begin{aligned}
&\lim_{i\to\infty} \frac{\#\{n \le N_i;\ \text{first } r \text{ digits of } p_n^s = k_1k_2\dots k_r\}}{N_i}\\
&= g_0\Big(\log_b\big(k_1.k_2k_3\dots(k_r+1)\big)\Big) - g_0\big(\log_b(k_1.k_2k_3\dots k_r)\big)\\
&= \frac{b^{(\log_b(k_1.k_2\dots(k_r+1)))/s}-1}{b^{1/s}-1} - \frac{b^{(\log_b(k_1.k_2\dots k_r))/s}-1}{b^{1/s}-1}\\
&= \frac{(k_1.k_2k_3\dots(k_r+1))^{(1/s)} - (k_1.k_2k_3\dots k_r))^{(1/s)}}{b^{1/s}-1}.
\end{aligned}$$

2.6.5. Two-dimensional Benford's law

Unsolved Problems [42, 1.38, p. 186]:

F. Luca and P. Stanica [23] proved

Theorem 23. *There exists infinite many n such that Fibonacci number F_n starts with digits K_1 and $\varphi(F_n)$ starts with digits K_2 in the base b representation. Here K_1 and K_2 are arbitrary and $\varphi(x)$ is the Euler function.*

In the following we see that this claim is equivalent that the sequence

$$\big(\log_b F_n, \log_b \varphi(F_n)\big) \bmod 1, \quad n = 1, 2, \ldots,$$

is everywhere dense in $[0,1]^2$, but the authors use the following method:

(i) By the first author $\varphi(F_n)/F_n$ is dense in $[0,1]$. Thus, for an interval I with arbitrary small length and containing K_2/K_1, there exists $\varphi(F_a)/F_a \in I$.

(ii) Then $\varphi(F_{ap})/F_{ap} \in I$ for all sufficiently large prime p.

(iii) There exists infinitely many primes p such that F_{ap} starts with K_1.

(iv) Finally, multiplying I by F_{ap} they find $\varphi(F_{ap})$ which starts with K_2.

Now we shall extend Theorem 20 to the two-dimensional case.

Let $x_n > 0$, $y_n > 0$, $n = 1, 2, \ldots$;

$F_N(x,y) = \frac{\#\{n \le N; \{\log_b x_n\} < x \text{ and } \{\log_b y_n\} < y\}}{N}$,

$K_1 = k_1^{(1)} k_2^{(1)} \ldots k_{r_1}^{(1)}$,

$K_2 = k_1^{(2)} k_2^{(2)} \ldots k_{r_2}^{(2)}$,

$u_1 = \log_b \big(\frac{K_1}{b^{r_1-1}}\big)$,

$u_2 = \log_b \big(\frac{K_1+1}{b^{r_1-1}}\big)$,

$v_1 = \log_b \big(\frac{K_2}{b^{r_2-1}}\big)$,

$v_2 = \log_b \big(\frac{K_2+1}{b^{r_2-1}}\big)$,

x_n has the first r_1 digits $= K_1 \iff \{\log_b x_n\} \in [u_1, u_2)$;

y_n has the first r_2 digits $= K_2 \iff \{\log_b y_n\} \in [v_1, v_2)$;

Theorem 24. *Let $g(x,y) \in G(\{\log_b x_n\}, \{\log_b y_n\})$ and $\lim_{k\to\infty} F_{N_k}(x,y) = g(x,y)$ for $(x,y) \in [0,1]^2$. Then*

$$\lim_{k\to\infty} \frac{\#\{n \le N_k; x_n \text{ has the first } r_1 \text{ digits } = K_1 \text{ and } y_n \text{ has the first } r_2 \text{ digits } = K_2\}}{N_k}$$
$$= g(u_2, v_2) + g(u_1, v_1) - g(u_2, v_1) - g(u_1, v_2). \tag{31}$$

Example 16.

$$G(\{\log_b n\}, \{\log_b (n+1)\})$$
$$= \left\{ g_u(x,y) = \frac{b^{\min(x,y)} - 1}{b-1} \frac{1}{b^u} + \frac{b^{\min(x,y,u)} - 1}{b^u}; u \in [0,1] \right\}.$$

$g_u(x,y) = \min(g_u(x), g_u(y))$ by Sklar theorem, where $g_u(x) = \frac{b^x - 1}{b-1} \cdot \frac{1}{b^u} + \frac{b^{\min(x,u)} - 1}{b^u}$. Thus

$$\lim_{k\to\infty} \frac{\#\{n \le N_k; n \text{ has the first } r_1 \text{ digits } = K_1 \text{ and } (n+1) \text{ has the first } r_2 \text{digits} = K_2\}}{N_k}$$
$$= g_u(u_2, v_2) + g_u(u_1, v_1) - g_u(u_2, v_1) - g_u(u_1, v_2)$$

If $K_1 = K_2$ then $= g_u(u_2) - g_u(u_1)$. It can be found directly.

Example 17. Let $x_n \in [0,1)$, $n = 1, 2, \ldots$, be a u.d. sequence. Then

(I) $$x_n \quad \text{and} \quad \log_b n \bmod 1$$
are statistically independent (Theorem 8);

(II) $$x_n \quad \text{and} \quad \log_b(n \log n) \bmod 1$$
are statistically independent (Y. Ohkubo (2011) [27]); see Theorem 9.

(III) $$x_n \quad \text{and} \quad \log_b p_n \bmod 1$$
are statistically independent (Y. Ohkubo (2011) [27]); see Theorem 10.

$$G(x_n, \{\log_b p_n\}) = \{x.g_u(y); u \in [0,1]\},$$

where $g_u(x) = \frac{b^x - 1}{b-1} \cdot \frac{1}{b^u} + \frac{b^{\min(x,u)} - 1}{b^u}$ and $F_{N_k}(x,y) \to x.g_u(y)$ if $\{\log_b N_k\} \to u$.

Example 18. Let $x_n \in [0,1)$, $n = 1, 2, \ldots$, be u.d. sequence. Then x_n and $\log_b n \bmod 1$ are statistically independent, i.e.,

$$G(x_n, \{\log_b n\}) = \{g_u(x,y) = x.g_u(y); u \in [0,1]\},$$

where $g_u(x) = \frac{b^x - 1}{b-1} \cdot \frac{1}{b^u} + \frac{b^{\min(x,u)} - 1}{b^u}$. This was proved by G. Rauzy (1973), see [44, p. 2–27, 2.3.6.]

Example 19. By Theorem 10 we have

$$G(\{\log_b F_n\}, \{\log_b p_n\}) = \{x.g_u(y); u \in [0,1]\}$$

and let

$$\{\log_b p_{N_k}\} \to u.$$

Then

$$\lim_{k\to\infty} \frac{\#\{n \le N_k; F_n \text{ has the first } r_1 \text{ digits} = K_1 \text{ and } p_n \text{ has the first } r_2 \text{ digits} = K_2\}}{N_k}$$
$$= u_2 g_u(v_2) + u_1 g_u(v_1) - u_2 g_u(v_1) - u_1 g_u(v_2), \tag{32}$$

where

$$u_1 = \log_b\left(\frac{K_1}{b^{r_1-1}}\right), \quad u_2 = \log_b\left(\frac{K_1+1}{b^{r_1-1}}\right),$$

$$v_1 = \log_b\left(\frac{K_2}{b^{r_2-1}}\right), \quad v_2 = \log_b\left(\frac{K_2+1}{b^{r_2-1}}\right),$$

and

$$g_u(x) = \frac{b^x - 1}{b-1} \cdot \frac{1}{b^u} + \frac{b^{\min(x,u)} - 1}{b^u}.$$

2.7. Two-dimensional copulas

[10] If distribution function $c(x,y)$ satisfies

$$c(x,1)=x \quad \text{and} \quad c(1,y)=y,$$

then $c(x,y)$ is called *copula*. These d.f.'s introduced by M. Sklar (1959) and all their basic properties can be found in R.B. Nelsen (1999).

Let $G_{2,1}$ be the set of all two-dimensional copulas. There are some basic properties $G_{2,1}$:

(I) $G_{2,1}$ is closed under pointwise limit and convex linear combinations.

(II) For every $g(x,y) \in G_{2,1}$ and every $(x_1,y_1),(x_2,y_2) \in [0,1]^2$ we have $|g(x_2,y_2)-g(x_1,y_1)| \le |x_2-x_1|+|y_2-y_1|$, [26, p. 9]. Also [26, p. 9, Coroll. 2.2.6]: The horizontal, vertical and diagonal sections of copula are all nondecreasing and uniformly continuous on $[0,1]$.

(III) For every $g(x,y) \in G_{2,1}$ we have $g_3(x,y)=\max(x+y-1,0) \le g(x,y) \le \min(x,y) = g_2(x,y)$, where $g_3(x,y)$ and $g_2(x,y)$ are copulas (Fréchet--Hoeffding bounds, see R.B. Nelsen (1999) [p. 9][26]).

(IV) M. Sklar (1959) proved (cf. (cf. Nelsen[p. 15, Th. 2.3.3])):

Theorem 25. *For every d.f. $g(x,y)$ on $[0,1]^2$ there exists $c(x,y) \in G_{2,1}$ such that*

$$g(x,y)=c(g(x,1),g(1,y)) \quad \textit{for every} \quad (x,y)\in[0,1]^2.$$

If $g(x,1)$ and $g(1,y)$ are continuous, then the copula $c(x,y)$ is unique .

• For d.f. $g(x,y)$ denote the marginal $g_1(x)=g(x,1)$ and $g_2(y)=g(1,y)$ and by Sklar $g(x,y)=c(g_1(x),g_2(y))$. Then for every continuous $F(x,y)$ we have

$$\int_0^1\int_0^1 F(x,y)\mathrm{d}g(x,y)=\int_0^1\int_0^1 F(g_1^{(-1)}(x),g_2^{(-1)}(y))\mathrm{d}c(x,y), \qquad (33)$$

see M. Hofer and M.R. Iacò [18].

(VI) Examples:

$g_\theta(x,y)=(\min(x,y))^\theta(xy)^{1-\theta}$, where $\theta\in[0,1]$ (Cuadras-Augé family, cf. Nelsen [1999, p. 12, Ex. 2.5]),

$g_4(x,y)=\frac{xy}{x+y-xy}$ (see Nelsen [1999, p. 19, 2.3.4]),

$\tilde{g}(x,y)=x+y-1+g(1-x,1-y)$ for every $g(x,y)\in G_{2,1}$ (Survival copula, see Nelsen [1999, p. 28, 2.6.1]).

[10] We denote by $G_{s,k}$ the set of all d.f.s $g(\mathbf{x})$ on $[0,1]^s$ for which all k-dimensional marginal (i.e., face) d.f.s satisfy $g(1,\dots,1,x_{i_1},1,\dots,1,x_{i_2},1,\dots,1,x_{i_k},1,\dots,1)=x_{i_1}x_{i_2}\dots x_{i_k}$. For $k=1$, these d.f.'s are called *copulas*

Possible a new types of copulas can be produced by: If a sequences (x_n, y_n) have both x_n and y_n u.d. then all d.f.'s $g(x, y)$ of the sequence (x_n, y_n) has marginals

$$g(x, 1) = x, \qquad g(1, y) = y.$$

Theorem 26. *Let X be a non-empty, closed and connected set of copulas. Then there exists a two-dimensional sequence (x_n, y_n), $n = 1, 2, \ldots$, in $[0, 1)^2$ such that $G(x_n, y_n) = X$.*

Example 20. For the sequence $(u + z_n, v + z_n) \bmod 1$, z_n, $n = 1, 2, \ldots$, z_n is u.d. we denote a.d.f. as $g_{u,v}(x, y)$. Then by Weyl's limit relation

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} K(\{u + z_n, v + z_n\}) = \int_0^1 K(\{u + z\}, \{v + z\})\mathrm{d}z$$
$$= \int_0^1 \int_0^1 K(x, y)\mathrm{d}_x\mathrm{d}_y g_{u,v}(x, y). \qquad (34)$$

In the following, for $g_{u,v}(x, y)$, we prove (39).

Let $0 \le u \le v \le 1$ be fixed and $x, y \in [0, 1]$ be variables and define

$$h_u(x) = x + u \bmod 1, \quad h_v(y) = y + v \bmod 1.$$

Then

$$g_{u,v}(x, y) = |h_u^{-1}([0, x)) \cap h_v^{-1}([0, y))|.$$

Using the following graphs of $h_u(x)$

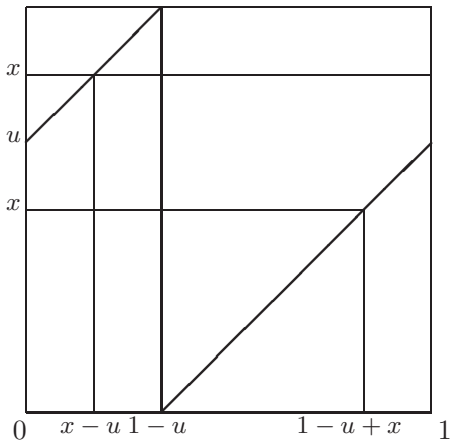


Figure 1: The graph of $h_u(x)$.

then we see

$$h_u^{-1}([0,x]) = \begin{cases} [1-u, 1-u+x] & \text{if } x \le u, \\ [0, x-u] \cup [1-u, 1] & \text{if } u \le x. \end{cases} \tag{35}$$

Hence

$$g_{u,v}(x,y) = \begin{cases} \big|[1-u,1-u+x] \cap [1-v,1-v+y]\big| & \text{if } x \le u, y \le v, \\ \big|[1-u,1-u+x] \cap ([0,y-v] \cup [1-v,1])\big| & \text{if } x \le u, y > v, \\ \big|([0,x-u] \cup [1-u,1]) \cap [1-v,1-v+y]\big| & \text{if } x > u, y \le v, \\ \big|([0,x-u] \cup [1-u,1]) \cap ([0,y-v] \cup [1-v,1])\big| & \text{if } x > u, y > v. \end{cases} \tag{36}$$

Now we used minimum and maximum formula for the length of intersection of two intervals $[\alpha,\beta]$ and $[\gamma,\delta]$

$$\big|[\alpha,\beta] \cap [\gamma,\delta]\big| = \max\big(\min(\beta,\delta) - \max(\alpha,\gamma), 0\big). \tag{37}$$

Insert (37) into (36) we see

$$g_{u,v}(x,y) = \begin{cases} \max\big(\min(y, x-u+v), 0\big) & \text{if } x \le u, y \le v, \\ \max\big(\min(x, y-v-1+u), 0\big) + \max(v-u+x, 0) & \text{if } x \le u, y > v, \\ y & \text{if } x > u, y \le v, \\ \min(x-u+v, y) + \max(y-v-1+u, 0) & \text{if } x > u, y > v, \end{cases} \tag{38}$$

which implies

$$g_{u,v}(x,y) = \begin{cases} x & \text{if } (x,y) \in A, \\ y - (1-|u-v|) & \text{if } (x,y) \in B, \\ x+y-1 & \text{if } (x,y) \in C, \\ 0 & \text{if } (x,y) \in D, \\ x - |u-v| & \text{if } (x,y) \in E, \\ y & \text{if } (x,y) \in F, \end{cases} \tag{39}$$

where

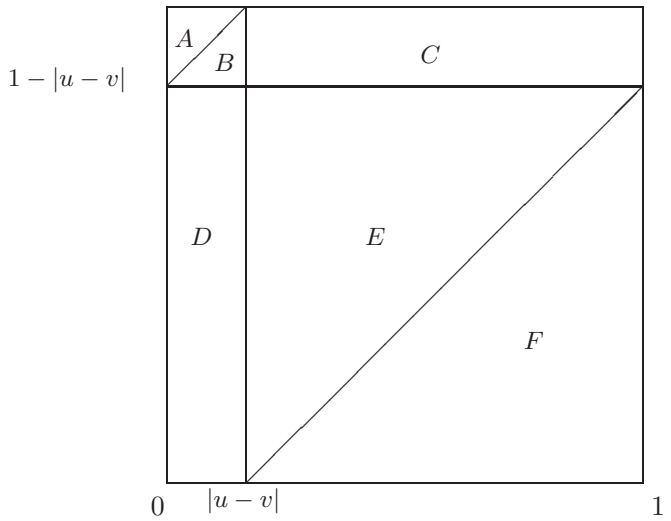


Figure 2: Division $[0,1]^2$ by (39).

2.7.1. Applications

Let $F(x,y)$ be a Riemann integrable function defined on $[0,1]^2$ and x_n, y_n, $n = 1,2,\ldots$, be two u.d. sequences in $[0,1)$. A problem is to find limit points of the sequence

$$\frac{1}{N}\sum_{n=1}^{N} F(x_n, y_n), \quad N = 1,2,\ldots \tag{40}$$

For $F(x,y) = |x-y|$, this problem was formulated by F. Pillichshammer and S. Steinerberger [30]. They proved:

Theorem 27. *Let x_n and y_n be two uniformly distributed sequences in $[0,1)$. Then*

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} |x_n - y_n| \le \frac{1}{2}$$

and in particular

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} |x_{n+1} - x_n| \le \frac{1}{2}$$

and this result is best possible.

They also found:

• $\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}|x_{n+1}-x_n| = \frac{2(b-1)}{b^2}$ for van der Corput sequence x_n in the base b and

• $\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}|x_{n+1}-x_n| = 2\{\alpha\}(1-\{\alpha\})$ for $x_n = n\alpha \bmod 1$, where α is irrational.

Applying Helly theorems we see that limit points of (40) form the set

$$\left\{\int_0^1\int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y g(x,y); g(x,y)\in G((x_n,y_n))\right\}, \tag{41}$$

where $G(x_n,y_n)$ is the set of all d.f.s of the two-dimensional sequence (x_n,y_n), $n=1,2,\dots$ In this case, two-dimensional sequence (x_n,y_n) does not need to be u.d. but every d.f. $g(x,y)\in G((x_n,y_n))$ satisfies:

(i) $g(x,1)=x$ for $x\in[0,1]$ and

(ii) $g(1,y)=y$ for $y\in[0,1]$.

Thus d.f. $g(x,y)$ is a copula.

2.8. Extremes of $\int_0^1\int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y g(x,y)$

The problem of to find extremes of (40) is equivalent to find extreme values of $\int_0^1\int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y g(x,y)$ over copulas $g(x,y)$. In [11] is proved:

Theorem 28. *Let $F(x,y)$ be a Riemann integrable function defined on $[0,1]^2$. For differential of $F(x,y)$ let us assume that $\mathrm{d}_x\mathrm{d}_y F(x,y) > 0$ for every $(x,y)\in(0,1)^2$. Then*

$$\max_{g(x,y)\text{-}copula}\int_0^1\int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y g(x,y) = \int_0^1 F(x,x)\mathrm{d}x,$$
$$\min_{g(x,y)\text{-}copula}\int_0^1\int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y g(x,y) = \int_0^1 F(x,1-x)\mathrm{d}x, \tag{42}$$

where, precisely, max *is attained in $g(x,y)=\min(x,y)$ and* min *in $g(x,y)=\max(x+y-1,0)$, uniquely.*

P r o o f. The integration by parts gives

$$\int_0^1\int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y g(x,y) = F(1,1) - \int_0^1 g(1,y)\mathrm{d}_y F(1,y) - \int_0^1 g(x,1)\mathrm{d}_x F(x,1) + \int_0^1\int_0^1 g(x,y)\mathrm{d}_x\mathrm{d}_y F(x,y) \tag{43}$$

which holds for every Riemann integrable $F(x,y)$ and d.f. $g(x,y)$ which does not have any common discontinuity points. Then Fréchet-Hoeffding bounds (see Nelsen [26, p. 9])

$$\max(x+y-1,0) \le g(x,y) \le \min(x,y)$$

and the assumption $\mathrm{d}_x\mathrm{d}_y F(x,y) > 0$ implies

$$\int_0^1\int_0^1 \max(x+y-1,0)\mathrm{d}_x\mathrm{d}_y F(x,y) \le \int_0^1\int_0^1 g(x,y)\mathrm{d}_x\mathrm{d}_y F(x,y) \le \int_0^1\int_0^1 \min(x,y)\mathrm{d}_x\mathrm{d}_y F(x,y).$$

Since every copula is continuous, then the left inequality is attained if and only if $g(x,y) = \max(x+y-1,0)$ and the right if and only if $g(x,y) = \min(x,y)$.

Directly by definition of a.d.f., for every u.d. sequence $x_n \in [0,1)$, it can be proved that

a) the sequence (x_n, x_n), $n = 1,2,\dots$, has the a.d.f. $g(x,y) = \min(x,y)$ and
b) the sequence $(x_n, 1-x_n)$, $n = 1,2,\dots$, has the a.d.f. $g(x,y) = \max(x+y-1,0)$. From it

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} F(x_n,x_n) = \int_0^1 F(x,x)\mathrm{d}x = \int_0^1\int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y \min(x,y), \tag{44}$$

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} F(x_n,1-x_n) = \int_0^1 F(x,1-x)\mathrm{d}x = \int_0^1\int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y \max(x+y-1,0). \tag{45}$$

If $\mathrm{d}_x\mathrm{d}_y F(x,y) < 0$, the right hand sides of (42) are exchanged.

□

Using copulas we gives in the following, an alternative proof of F. Pillichshammer and S. Steinerberger [30] Theorem 27:

Example 21. Putting $F(x,y) = |x-y|$, we have $F(1,1) = 0$, $F(1,x) = 1-x$, $F(y,1) = 1-y$, and computing, for $y > x$,

$$\mathrm{d}_x\mathrm{d}_y|y-x| = (y+\mathrm{d}y-(x+\mathrm{d}x)) = (y-x)-(y-(x+\mathrm{d}x))-(y+\mathrm{d}y-x) = 0,$$

and for $y = x$, $\mathrm{d}y = \mathrm{d}x$,

$$\mathrm{d}_x\mathrm{d}_y|y-x| = |x+\mathrm{d}x-(x+\mathrm{d}x)|+|x-x|-|(x+\mathrm{d}x)-x|-|x-(x+\mathrm{d}x)| = -2\mathrm{d}x$$

then we have

$$\int_0^1 \int_0^1 |x-y| \mathrm{d}_x \mathrm{d}_y g(x,y) = \int_0^1 g(x,1)\mathrm{d}x + \int_0^1 g(1,y)\mathrm{d}y - 2\int_0^1 g(x,x)\mathrm{d}x. \quad (46)$$

Thus for a copula $g(x,y)$, $g(x,1)=x$, $g(1,y)=y$ we have

$$\int_0^1 \int_0^1 |x-y| \mathrm{d}_x \mathrm{d}_y g(x,y) = 1 - 2\int_0^1 g(x,x)\mathrm{d}x. \quad (47)$$

We shall compute (47) for van der Corput sequence $\gamma_q(n)$, $n=0,1,\ldots$, in base q. We have that every point $(\gamma_q(n), \gamma_q(n+1))$, $n=0,1,2,\ldots$, lies on the diagonals of intervals

$$\left[0, 1-\frac{1}{q}\right] \times \left[\frac{1}{q}, 1\right] \quad (48)$$

$$\left[1-\frac{1}{q^i}, 1-\frac{1}{q^{i+1}}\right] \times \left[\frac{1}{q^{i+1}}, \frac{1}{q^i}\right], \qquad i=1,2,\ldots \quad (49)$$

By Fig. 1 we find the so-called von Neumann-Kakutani transformation $T : [0,1] \to [0,1]$, see Fig. 1. Because $\gamma_q(n)$ is u.d., the sequence $\big(\gamma_q(n), \gamma_q(n+1)\big)$ has a.d.f. $g(x,y)$ which is copula of the form

$$\begin{aligned} g(x,y) =& |\mathrm{Project}_X(([0,x]\times[0,y]) \cap \text{graph } T)| \\ =& \min\big(\big|[0,x]\cap I_X\big|, \big|[0,y]\cap I_Y\big|\big) \\ &+ \sum_{i=1}^{\infty} \min\big(\big|[0,x]\cap I_X^{(i)}\big|, \big|[0,y]\cap I_Y^{(i)}\big|\big), \end{aligned} \quad (50)$$

where $\mathrm{Project}_X$ is the projection of a two dimensional set to the X-axis. [11]

[11] Copula $g(x,y)$ of the type (50) is called *shuffle of M* (see [26, p. 69]). It is a copula whose support is a collection of line segments with slope $+1$ or -1.

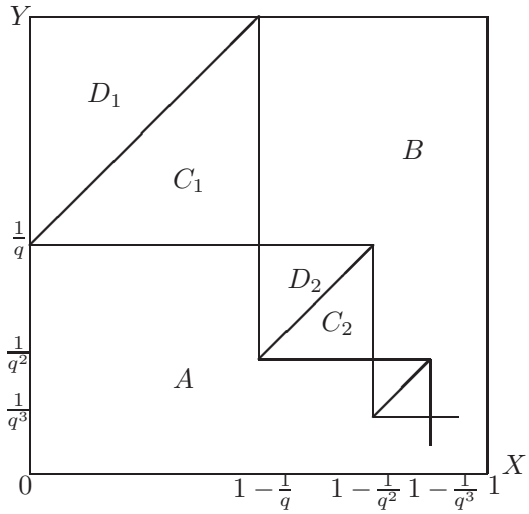


Figure 1: Line segments containing $(\gamma_q(n), \gamma_q(n+1))$, $n = 1, 2, \ldots$
The graph of the von Neumann-Kakutani transformation T.

The sum (50) implies

$$g(x,y) = \begin{cases} 0 & \text{if } (x,y) \in A, \\ 1-(1-y)-(1-x) = x+y-1 & \text{if } (x,y) \in B, \\ y - \frac{1}{q^i} & \text{if } (x,y) \in C_i, \\ x - 1 + \frac{1}{q^{i-1}} & \text{if } (x,y) \in D_i, \end{cases} \tag{51}$$

$i = 1, 2, \ldots$ From (51) it follows

$$g(x,x) = \begin{cases} 0, & \text{if } x \in \left[0, \frac{1}{q}\right], \\ x - \frac{1}{q}, & \text{if } x \in \left[\frac{1}{q}, 1 - \frac{1}{q}\right], \\ 2x - 1, & \text{if } x \in \left[1 - \frac{1}{q}, 1\right], \end{cases} \tag{52}$$

(for $q = 2$, the mean equality misses) and by (47)

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} |\gamma_q(n) - \gamma_q(n+1)| = 1 - 2\int_0^1 g(x,x)\mathrm{d}x = \frac{2(q-1)}{q^2}.$$

As we see in part 2.7.1 in uniform distribution theory the problem of optimizing the integral

$$\int_0^1 \int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y g(x,y) \tag{53}$$

over copulas $g(x,y)$ is motivated by computing optimal limit points of the sequence $\frac{1}{N}\sum_{n=1}^{N} F(x_n,y_n)$, $N=1,2,\dots$ over uniform distribution sequences x_n and y_n, $n=1,2,\dots$ But problem of optimizing (53) belongs to the well-known mass transportation problems, or the Monge-Kantorovich transportation problem, see e.g., L. Ambrosio and N. Gigli (2013) [1]. In Theorem 28 we have seen that the solution of the problem in uniform distribution theory depends on the sign of partial derivatives $\frac{\partial^2 F(x,y)}{\partial x\partial y}$, see Fig. 2.

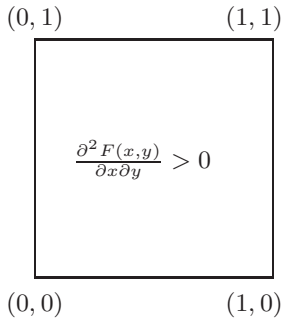


Figure 2.

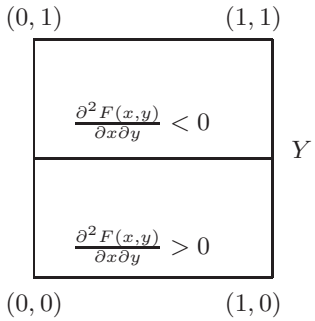


Figure 3.

A criterion for Fig. 3 follows from [11, Th. 7]:

Theorem 29. *Let us assume that a copula $g(x,y)$ maximizes the integral*

$$\int_0^1\int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y g(x,y).$$

Let $[X_1,X_2]\times[Y_1,Y_2]$ be an interval in $[0,1]^2$ such that the differential

$$g(X_2,Y_2)+g(X_1,Y_1)-g(X_2,Y_1)-g(X_1,Y_2)>0.$$

Assume that for every interior point (x,y) of $[X_1,X_2]\times[Y_1,Y_2]$ the differential $\mathrm{d}_x\mathrm{d}_y F(x,y)$ has constant signum. Then we have:

(i) *if $\mathrm{d}_x\mathrm{d}_y F(x,y)>0$, then*

$$g(x,y)=\min\big(g(x,Y_2)+g(X_1,y)-g(X_1,Y_2),\\ g(x,Y_1)+g(X_2,y)-g(X_2,Y_1)\big) \quad (54)$$

(ii) *if $\mathrm{d}_x\mathrm{d}_y F(x,y)<0$, then*

$$g(x,y)=\max\big(g(x,Y_2)+g(X_2,y)-g(X_2,Y_2),\\ g(x,Y_1)+g(X_1,y)-g(X_1,Y_1)\big) \quad (55)$$

for every $(x,y)\in[X_1,X_2]\times[Y_1,Y_2]$.

Theorem 29 can also be used for Fig. 4. Such a method is described in R.F. Tichy, S. Thonhauser, O. Strauch, M.R. Iacó and V. Baláž [49]:

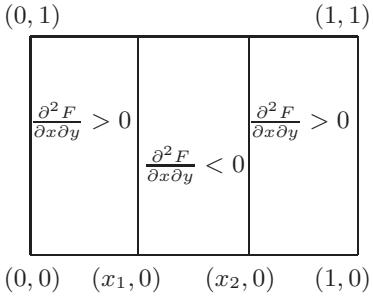


Figure 4.

Let $g(x,y)$ be a copula which maximize $\int_0^1 \int_0^1 F(x,y) \mathrm{d}_x \mathrm{d}_y g(x,y)$ and $F(x,y)$ satisfies Fig. 4. Then $g(x,y)$ satisfies (54) if $x \in (0,x_1) \cup (x_2,1)$ and (55) if $x \in (x_1,x_2)$. Denote

$$g(x_1,y) = h_1(y), \quad \text{and} \quad g(x_2,y) = h_2(y).$$

Thus:
If $x \in (0,x_1)$, then

$$\begin{aligned} g(x,y) &= \min(g(0,y) + g(x,1) - g(0,1), g(x,0) + g(x_1,y) - g(x_1,0) \\ &= \min(0 + x - 0, 0 + h_1(y) - 0) \\ &= \min(x, h_1(y)). \end{aligned}$$

If $x \in (x_1,x_2)$, then

$$\begin{aligned} g(x,y) &= \max(g(x,1) + g(x_2,y) - g(x_2,1), g(x_1,y) + g(x,0) - g(x_1,0) \\ &= \max(x + h_2(y) - x_2, h_1(y) + 0 - 0) \\ &= \max(x + h_2(y) - x_2, h_1(y)). \end{aligned}$$

If $x \in (x_2,1)$, then

$$\begin{aligned} g(x,y) &= \min(g(x_2,y) + g(x,1) - g(x_2,1), g(x,0) + g(1,y) - g(1,0) \\ &= \min(h_2(y) + x - x_2, 0 + y - 0) \\ &= \min(x - x_2 + h_2(y), y). \end{aligned}$$

Summary

$$g(x,y)=\begin{cases}\min\bigl(x,h_1(y)\bigr) & \text{if } x\in[0,x_1],\\ \max\bigl(x+h_2(y)-x_2,h_1(y)\bigr) & \text{if } x\in[x_1,x_2],\\ \min\bigl(x-x_2+h_2(y),y\bigr) & \text{if } x\in[x_2,1]\end{cases}\tag{56}$$

where $y\in[0,1]$ and $h_1(y)$ and $h_2(y)$ we must calculated.
The differential $\mathrm{d}_x\mathrm{d}_y g(x,y)$ is nonzero only for points (x,y) on the curves

$$\begin{aligned}&x=h_1(y), y\in[0,1],\\ &x=x_2-h_2(y)+h_1(y), y\in[0,1],\\ &x=x_2-h_2(y)+y, y\in[0,1],\end{aligned}\tag{57}$$

and we have

$$\mathrm{d}_x\mathrm{d}_y g(x,y)=\begin{cases}h_1'(y)\mathrm{d}y & \text{if } x\in[0,x_1], x=h_1(y),\\ (h_2'(y)-h_1'(y))\mathrm{d}y & \text{if } x\in[x_1,x_2], x=x_2-h_2(y)+h_1(y),\\ (1-h_2'(y))\mathrm{d}y & \text{if } x\in[x_2,1], x=x_2-h_2(y)+y.\end{cases}\tag{58}$$

Let

$$F(x,y)=\begin{cases}F_1(x,y) & \text{if } x\in(0,x_1), \frac{\partial^2 F_1(x,y)}{\partial x\partial y}>0\\ F_2(x,y) & \text{if } x\in(x_1,x_2), \frac{\partial^2 F_2(x,y)}{\partial x\partial y}<0\\ F_3(x,y) & \text{if } x\in(x_2,1), \frac{\partial^2 F_3(x,y)}{\partial x\partial y}>0.\end{cases}$$

Then (56), (58) and (57) give

$$\begin{aligned}\int_0^1\int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y g(x,y)&=\int_0^1 F_1\bigl(h_1(y),y\bigr)h_1'(y)\mathrm{d}y\\ &\quad+\int_0^1 F_2\bigl(x_2-h_2(y)+h_1(y),y\bigr)\bigl(h_2'(y)-h_1'(y)\bigr)\mathrm{d}y\\ &\quad+\int_0^1 F_3\bigl(x_2-h_2(y)+y,y\bigr)\bigl(1-h_2'(y)\bigr)\mathrm{d}y.\end{aligned}\tag{59}$$

Denote

$$\begin{aligned}&G(y,h_1,h_2,h_1',h_2')\\ &=F_1\bigl(h_1(y),y\bigr)h_1'(y)+F_2\bigl(x_2-h_2(y)+h_1(y),y\bigr)\bigl(h_2'(y)-h_1'(y)\bigr)\\ &+F_3\bigl(x_2-h_2(y)+y,y\bigr)(1-h_2'(y)).\end{aligned}$$

Then

$$\max_{g(x,y)-\text{ copula}} \int_0^1 \int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y g(x,y) = \max_{h_1,h_2} \int_0^1 G(y,h_1,h_2,h_1',h_2')\mathrm{d}y, \quad (60)$$

where h_1, h_2 give a copula in (56). To do this we use the following criterion:

Theorem 30. *The function $g(x,y)$ defined by* (56) *is a copula if and only if*

(i) $h_1(y)$ *and* $h_2(y)$ *are increasing;*

(ii) $h_1(0) = 0$, $h_2(0) = 0$*;*

(iii) $h_1(1) = x_1$, $h_2(1) = x_2$*;*

(iv) $0 \le h_1(y) \le h_2(y) \le y$*;*

(v) $0 \le h_1'(y) \le h_2'(y) \le 1$*;*

If the (h_1, h_2) is maximum of $\int_0^1 G(y,h_1,h_2,h_1',h_2')\mathrm{d}y$ but not satisfy assumptions of Theorem 30, then we have only

$$\max_{g(x,y)-\text{copula}} \int_0^1 \int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y g(x,y) \le \int_0^1 G(y,h_1,h_2,h_1',h_2')\mathrm{d}y. \quad (61)$$

For solution of (60) we can using the calculus of variations: If (h_1, h_2) extremize the integral $\int_0^1 G(y,h_1,h_2,h_1',h_2')\mathrm{d}y$ then (h_1, h_2) must be satisfied the following system of Euler-Lagrange differential equations

$$\frac{\partial G}{\partial h_1} - \frac{\mathrm{d}}{\mathrm{d}y}\frac{\partial G}{\partial h_1'} = 0,$$
$$\frac{\partial G}{\partial h_2} - \frac{\mathrm{d}}{\mathrm{d}y}\frac{\partial G}{\partial h_2'} = 0.$$

The solution (h_1, h_2) maximize $\int_0^1 G(y,h_1,h_2,h_1',h_2')\mathrm{d}y$ if

$$\frac{\partial^2 G}{\partial h_1'\partial h_1'} \le 0, \quad \begin{vmatrix} \frac{\partial^2 G}{\partial h_1'\partial h_1'} & \frac{\partial^2 G}{\partial h_1'\partial h_2'} \\ \frac{\partial^2 G}{\partial h_2'\partial h_1'} & \frac{\partial^2 G}{\partial h_2'\partial h_2'} \end{vmatrix} \le 0.$$

To compare (60)we give L. Uckelmann's (1997) [48] the mass transportation problems: Let

$F(x,y) = \Phi(x+y)$ for $(x,y) \in [0,1]^2$;

For $0 < k_1 < k_2 < 2$ let $\Phi(x)$ be a twice differentiable function such that $\Phi(x)$ is strictly convex on $[0,k_1] \cup [k_2,2]$ and concave on $[k_1,k_2]$, i.e.,

$\Phi''(x) > 0$ for $x \in [0,k_1) \cup (k_2,2]$',

$\Phi''(x) > 0$ for $x \in [0,k_1) \cup (k_2,2]$.

If α and β are the solutions of

$$\Phi(2\alpha) - \Phi(\alpha+\beta) + (\beta-\alpha)\Phi'(\alpha+\beta) = 0,$$
$$\Phi(2\beta) - \Phi(\alpha+\beta) + (\alpha-\beta)\Phi'(\alpha+\beta) = 0$$

such that $0 < \alpha < \beta < 1$, then the optimal copula $C(x,y)$ is the shuffle of M

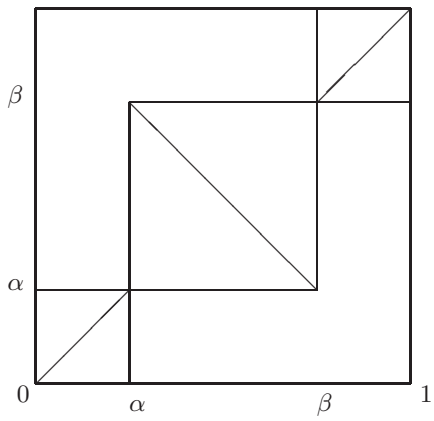


Figure 5.

with the support

$$\Gamma(x) = \begin{cases} x & \text{for } x \in [0,\alpha] \cup [\beta,1], \\ \alpha+\beta-x & \text{for } x \in (\alpha,\beta). \end{cases}$$

Then

$$\max \int_0^1 \int_0^1 F(x,y)dC(x,y) = \int_0^\alpha \Phi(2x)dx + (\beta-\alpha)\Phi(\alpha+\beta) + \int_\beta^1 \Phi(2x)dx.$$

2.9. Example of three-dimensional copula

See [12]: In this part we apply the Weyl's limit relation to calculate the limit

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} F\big(\gamma_q(n), \gamma_q(n+1), \gamma_q(n+2)\big)$$
$$= \int_0^1 \int_0^1 \int_0^1 F(x,y,z)\mathrm{d}_x\mathrm{d}_y\mathrm{d}_z g(x,y,z),$$

where $\gamma_q(n)$ is the van der Corput sequence in base q, $g(x,y,z)$ is the asymptotic distribution function of $\big(\gamma_q(n), \gamma_q(n+1), \gamma_q(n+2)\big)$, and $F(x,y,z) = \max(x,y,z)$.

Let $q \geq 3$ be an integer.

We start with that every point $\big(\gamma_q(n), \gamma_q(n+1)\big)$, $n = 0, 1, 2, \ldots$, lies on the diagonals of intervals (48) and (49). Then all terms of the sequence $\big(\gamma_q(n), \gamma_q(n+2)\big), n = 0, 1, 2, \ldots$, lie in the diagonals of the following intervals

$$\left[0, 1 - \frac{2}{q}\right] \times \left[\frac{2}{q}, 1\right], \tag{62}$$

$$\left[1 - \frac{1}{q^i}, 1 - \frac{1}{q^{i+1}}\right] \times \left[\frac{1}{q} + \frac{1}{q^{i+1}}, \frac{1}{q} + \frac{1}{q^i}\right], i = 1, 2, \ldots, \tag{63}$$

$$\left[1 - \frac{1}{q} - \frac{1}{q^k}, 1 - \frac{1}{q} - \frac{1}{q^{k+1}}\right] \times \left[\frac{1}{q^{k+1}}, \frac{1}{q^k}\right], k = 1, 2, \ldots, \tag{64}$$

Every maximal 3-dimensional interval I containing points

$$\big(\gamma_q(n),\, \gamma_q(n+1),\, \gamma_q(n+2)\big)$$

will be written as $I = I_X \times I_Y \times I_Z$, where I_X, I_Y, I_Z are projections of I to the X, Y, Z, axes, respectively. Moreover if

$$\gamma_q(n) \in I_X, \quad \text{then} \quad \gamma_q(n+1) \in I_Y \quad \text{and} \quad \gamma_q(n+2) \in I_Z.$$

From u.d. of $\gamma_q(n)$ follows that the lengths $|I_X| = |I_Y| = |I_Z|$. Combining intervals (48), (62), (63), (64), (49) of equal lengths by following Figure 3, then we find that every point $\big(\gamma_q(n), \gamma_q(n+1), \gamma_q(n+2)\big)$ is contained in diagonals of the intervals

$$I = \left[0, 1 - \frac{2}{q}\right] \times \left[\frac{1}{q}, 1 - \frac{1}{q}\right] \times \left[\frac{2}{q}, 1\right], \tag{65}$$

$$I^{(i)} = \left[1 - \frac{1}{q^i}, 1 - \frac{1}{q^{i+1}}\right] \times \left[\frac{1}{q^{i+1}}, \frac{1}{q^i}\right] \times \left[\frac{1}{q} + \frac{1}{q^{i+1}}, \frac{1}{q} + \frac{1}{q^i}\right], i = 1, 2, \ldots, \tag{66}$$

$$J^{(k)} = \left[1 - \frac{1}{q} - \frac{1}{q^k}, 1 - \frac{1}{q} - \frac{1}{q^{k+1}}\right] \times \left[1 - \frac{1}{q^k}, 1 - \frac{1}{q^{k+1}}\right] \times \left[\frac{1}{q^{k+1}}, \frac{1}{q^k}\right], \quad k = 1, 2, \ldots, \tag{67}$$

where $|I| = 0$ if $q = 2$. These intervals are maximal with respect to inclusion, see Fig. 1:

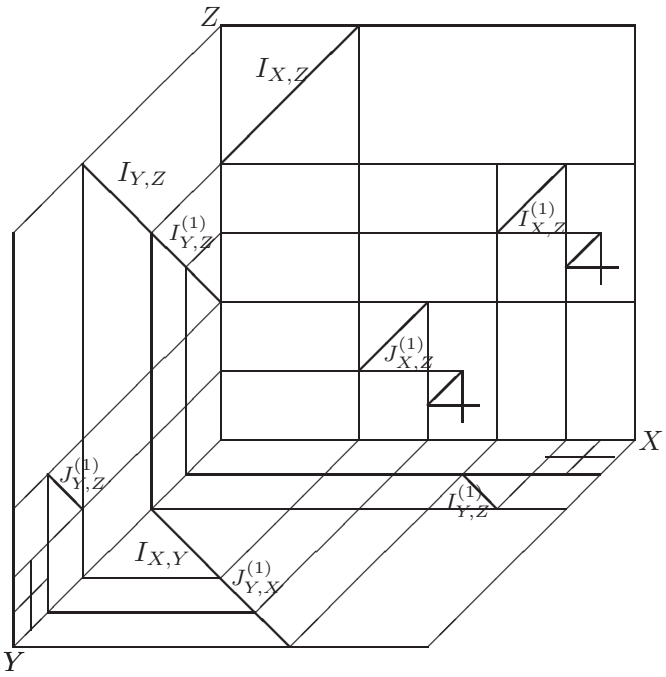


Figure 1: Mapping between intervals with equal lengths.

Now, let T be the union of diagonals of (66), (67) and (65). Again, as in (50), the a.d.f. $g(x,y,z)$ has the form [12]

$$g(x,y,z) = |\mathrm{Project}_X([0,x]\times[0,y]\times[0,z]\cap T)| \tag{68}$$

and it can be rewritten as

$$\begin{aligned} g(x,y,z) &= \min\left(|[0,x]\cap I_X|, |[0,y]\cap I_Y|, |[0,z]\cap I_Z|\right) \\ &\quad + \sum_{i=1}^{\infty} \min\left(\left|[0,x]\cap I_X^{(i)}\right|, \left|[0,y]\cap I_Y^{(i)}\right|, \left|[0,z]\cap I_Z^{(i)}\right|\right) \\ &\quad + \sum_{k=1}^{\infty} \min\left(\left|[0,x]\cap J_X^{(k)}\right|, \left|[0,y]\cap J_Y^{(k)}\right|, \left|[0,z]\cap J_Z^{(k)}\right|\right). \end{aligned} \tag{69}$$

To calculate minimums in (69) we can use the following Fig. 2 (here $q=3$):

[12] Since $g(x,y,z)$ is continuous, we use in the calculation closed intervals.

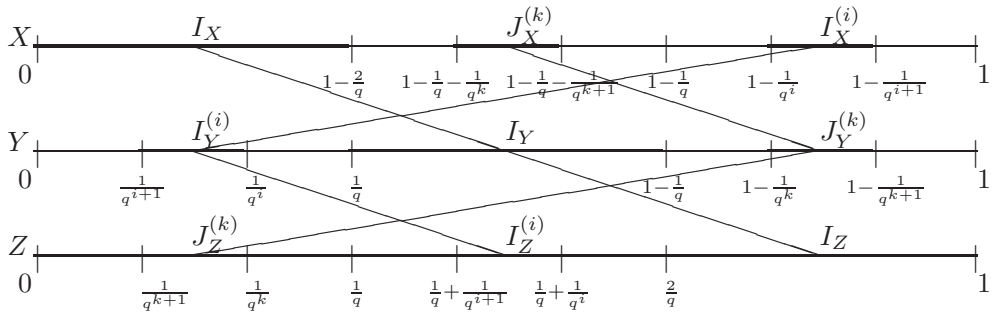


Figure 2: Projections of intervals $I, I^{(i)}, J^{(k)}$ on axes X, Y, Z.

As an example of application of (69) and Fig. 2 we compute $g(x,x,x)$ for $q \geq 3$ without the knowledge of $g(x,y,z)$, [13]

$$g(x,x,x) = \begin{cases} 0 & \text{if } x \in \left[0, \frac{2}{q}\right], \\ x - \frac{2}{q} & \text{if } x \in \left[\frac{2}{q}, 1 - \frac{1}{q}\right], \\ 3x - 2 & \text{if } x \in \left[1 - \frac{1}{q}, 1\right]. \end{cases} \tag{70}$$

P r o o f.

1. Let $x \in \left[0, \frac{1}{q}\right]$.
Then $|[0,x] \cap I_Z| = 0$, $\left|[0,x] \cap I_Z^{(i)}\right| = 0$, $\left|[0,x] \cap J_Y^{(k)}\right| = 0$, consequently $g(x,x,x) = 0$.

2. Let $x \in \left[\frac{1}{q}, \frac{2}{q}\right]$.
Then $|[0,x] \cap I_Z| = 0$, $\left|[0,x] \cap J_Y^{(k)}\right| = 0$, $\left|[0,x] \cap I_X^{(i)}\right| = 0$, consequently $g(x,x,x) = 0$.

3. Let $x \in \left[\frac{2}{q}, 1 - \frac{1}{q}\right]$.
Then $\left|[0,x] \cap I_X^{(i)}\right| = 0$, $\left|[0,x] \cap J_Y^{(k)}\right| = 0$, consequently
$g(x,x,x) = \min\left(1 - \frac{2}{q}, x - \frac{1}{q}, x - \frac{2}{q}\right) = x - \frac{2}{q}$.

4. Let $x \in \left[1 - \frac{1}{q}, 1\right]$.
Specify $x \in I_X^{(k_1)}$, $x \in J_Y^{(k_1)}$. Then $\left|[0,x] \cap I_X^{(k)}\right| = 0$, $\left|[0,x] \cap J_Y^{(k)}\right| = 0$ for $k > k_1$. Thus (69) implies

[13] For $q = 3$ the middle member in (70) is omitted.

$$
\begin{aligned}
g(x,x,x) &= \min\left(1-\frac{2}{q}, 1-\frac{1}{q}-\frac{1}{q}, x-\frac{2}{q}\right)\\
&\quad + \sum_{i=1}^{k_1} \min\left(\left|[0,x]\cap I_X^{(i)}\right|, \left|[0,y]\cap I_Y^{(i)}\right|, \left|[0,z]\cap I_Z^{(i)}\right|\right)\\
&\quad + \sum_{k=1}^{k_1} \min\left(\left|[0,x]\cap J_X^{(k)}\right|, \left|[0,y]\cap J_Y^{(k)}\right|, \left|[0,z]\cap J_Z^{(k)}\right|\right)\\
&= x-\frac{2}{q} + \sum_{i=1}^{k_1-1}\left(\frac{1}{q^i}-\frac{1}{q^{i+1}}\right) + x - 1 + \frac{1}{q^{k_1}}\\
&\quad + \sum_{k=1}^{k_1-1}\left(\frac{1}{q^k}-\frac{1}{q^{k+1}}\right) + x - 1 + \frac{1}{q^{k_1}} = 3x-2.
\end{aligned}
$$

□

For $q=2$ we have

$$
g(x,x,x) = \begin{cases} 0 & \text{if } x\in\left[0,\frac{1}{2}\right],\\ x-\frac{1}{2} & \text{if } x\in\left[\frac{1}{2},\frac{3}{4}\right],\\ 3x-2 & \text{if } x\in\left[\frac{3}{4},1\right]. \end{cases} \tag{71}
$$

The knowledge [14] of the a.d.f. $g(x,y,z)$ of the sequence

$$
\big(\gamma_q(n),\gamma_q(n+1),\gamma_q(n+2)\big), n=1,2,\ldots
$$

allows us to compute the following limit by the Weyl limit relation (4) in dimension $s=3$.

$$
\begin{aligned}
&\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} F\big(\gamma_q(n),\gamma_q(n+1),\gamma_q(n+2)\big)\\
&= \int_0^1\int_0^1\int_0^1 F(x,y,z)\mathrm{d}_x\mathrm{d}_y\mathrm{d}_x g(x,y,z),
\end{aligned} \tag{72}
$$

where $F(x,y,z)$ is an arbitrary continuous function defined in $[0,1]^3$. To calculate the Riemann-Stieltjes integral (72) we will use the integration by parts.

Lemma 31. *Assume that $F(x,y,z)$ is a continuous in $[0,1]^3$ and $g(x,y,z)$ is a d.f.*

[14] $g(x,y,z)$ is explicitly given in [12] by using 27 possibilities.

Then

$$\begin{aligned}
&\int_0^1\int_0^1\int_0^1 F(x,y,z)\mathrm{d}_x\mathrm{d}_y\mathrm{d}_z g(x,y,z)\\
&= F(1,1,1,) - \int_0^1 g(1,1,z)\mathrm{d}_z F(1,1,z) - \int_0^1 g(1,y,1)\mathrm{d}_y F(1,y,1)\\
&\quad - \int_0^1 g(x,1,1)\mathrm{d}_x F(x,1,1) + \int_0^1\int_0^1 g(1,y,z)\mathrm{d}_y\mathrm{d}_z F(1,y,z)\\
&\quad + \int_0^1\int_0^1 g(x,1,z)\mathrm{d}_x\mathrm{d}_z F(x,1,z) + \int_0^1\int_0^1 g(x,y,1)\mathrm{d}_x\mathrm{d}_y F(x,y,1)\\
&\quad - \int_0^1\int_0^1\int_0^1 g(x,y,z)\mathrm{d}_x\mathrm{d}_y\mathrm{d}_z F(x,y,z). \qquad (73)
\end{aligned}$$

Here

$$\begin{aligned}
\mathrm{d}_x\mathrm{d}_y F(x,y) &= F(x+\mathrm{d}x, y+\mathrm{d}y) + F(x,y) - F(x+\mathrm{d}x,y) - F(x,y+\mathrm{d}y),\\
\mathrm{d}_x\mathrm{d}_y\mathrm{d}_z F(x,y,z) &= F(x+\mathrm{d}x, y+\mathrm{d}y, z+\mathrm{d}z) - F(x,y,z)\\
&\quad + F(x+\mathrm{d}x,y,z) + F(x,y+\mathrm{d}y,z) + F(x,y,z+\mathrm{d}z)\\
&\quad - F(x+\mathrm{d}x,y+\mathrm{d}y,z) - F(x,y+\mathrm{d}y,z+\mathrm{d}z)\\
&\quad - F(x+\mathrm{d}x,y,z+\mathrm{d}z). \qquad (74)
\end{aligned}$$

Note that

$$\mathrm{d}_x\mathrm{d}_y F(x,y) = \frac{\partial^2 F(x,y)}{\partial x\partial y}\mathrm{d}x\mathrm{d}y,$$

$$\mathrm{d}_x\mathrm{d}_y\mathrm{d}_z F(x,y,z) = \frac{\partial^3 F(x,y,z)}{\partial x\partial y\partial z}\mathrm{d}x\mathrm{d}y\mathrm{d}z,$$

if the partial derivatives exist. Put

$$F(x,y,z) = \max(x,y,z).$$

We have

$$\mathrm{d}_x F(x,1,1) = \mathrm{d}_y F(1,y,1) = \mathrm{d}_z F(1,1,z) = 0,$$

$$\mathrm{d}_x\mathrm{d}_y F(x,y,1) = \mathrm{d}_x\mathrm{d}_z F(x,1,z) = \mathrm{d}_y\mathrm{d}_z F(1,y,z) = 0,$$

The differential $\mathrm{d}_x\mathrm{d}_y\mathrm{d}_z F(x,y,z)$ is non-zero if and only if $x = y = z$ and in this case $\mathrm{d}_x\mathrm{d}_y\mathrm{d}_z F(x,y,z) = \mathrm{d}x$.

P r o o f. For every interval

$$J = \left[x_1^{(1)}, x_2^{(1)}\right] \times \left[x_1^{(2)}, x_2^{(2)}\right] \times \cdots \times \left[x_1^{(s)}, x_2^{(s)}\right] \subset [0,1]^s$$

and every continuous $F(x_1, x_2, \ldots, x_s)$ the differential $\Delta(F, J)$ is defined as

$$\Delta(F,J) = \sum_{\varepsilon_1=1}^{2} \cdots \sum_{\varepsilon_s=1}^{2} (-1)^{\varepsilon_1 + \cdots + \varepsilon_s} F(x_{\varepsilon_1}^{(1)}, \ldots, x_{\varepsilon_s}^{(s)}). \tag{75}$$

Putting $F(x_1, x_2, \ldots, x_s) = \max(x_1, x_2, \ldots, x_s)$, $x_1^{(i)} = x$, $x_2^{(i)} = x + \mathrm{d}x$ we have

$$\begin{aligned}
\Delta(F,J) &= (-1)^{1+1+\cdots+1} x + \sum_{\varepsilon_1=1}^{2} \cdots \sum_{\varepsilon_s=1}^{2} (-1)^{\varepsilon_1+\cdots+\varepsilon_s} (x + \mathrm{d}x) \\
&= \sum_{\varepsilon_1=1}^{2} \cdots \sum_{\varepsilon_s=1}^{2} (-1)^{\varepsilon_1+\cdots+\varepsilon_s} (x + \mathrm{d}x) - (-1)^{1+1+\cdots+1} \mathrm{d}x \\
&= (-1)^{s+1} \mathrm{d}x.
\end{aligned}$$

□

Then by (73)

$$\begin{aligned}
&\int_0^1 \int_0^1 \int_0^1 F(x,y,z) \mathrm{d}_x \mathrm{d}_y \mathrm{d}_z g(x,y,z) \\
&= 1 - \int_0^1 \int_0^1 \int_0^1 g(x,y,z) \mathrm{d}_x \mathrm{d}_y \mathrm{d}_z F(x,y,z) \\
&= 1 - \int_0^1 g(x,x,x) \mathrm{d}x.
\end{aligned} \tag{76}$$

For $q \geq 3$ and by (70) we have

$$\int_0^1 g(x,x,x)\mathrm{d}x = \int_{\frac{2}{q}}^{1-\frac{1}{q}} \left(x - \frac{2}{q}\right) \mathrm{d}x + \int_{1-\frac{1}{q}}^{1} (3x-2)\mathrm{d}x = \frac{1}{2} - \frac{2}{q} + \frac{3}{q^2}.$$

For $q = 2$ and by (71) we have

$$\int_0^1 g(x,x,x)\mathrm{d}x = \int_{\frac{1}{2}}^{\frac{3}{4}} \left(x - \frac{1}{2}\right) \mathrm{d}x + \int_{\frac{3}{4}}^{1} (3x-2)\mathrm{d}x = \frac{3}{16}.$$

Therefore for $q \geq 3$, by (72) and by (76) we have

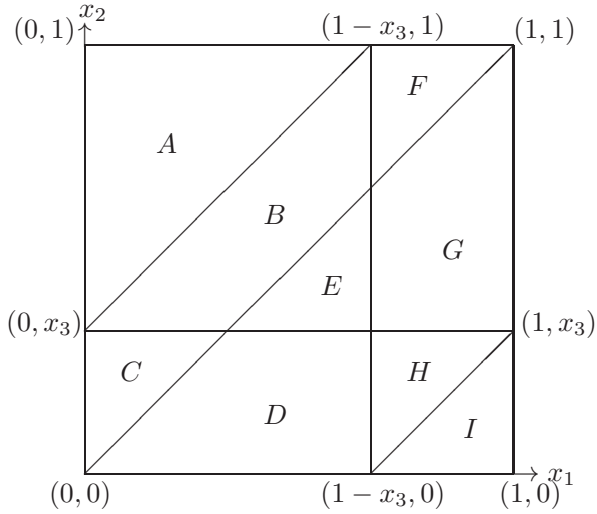
$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \max\bigl(\gamma_q(n), \gamma_q(n+1), \gamma_q(n+2)\bigr) = \frac{1}{2} + \frac{2}{q} - \frac{3}{q^2}. \tag{77}$$

Example 22. Example of a copula in $G_{3,2}$ [44, 3-5,3.2.8].

Let u_n and v_n be two u.d. and statistically independent sequences in $[0,1)$. Then the sequence

$$\mathbf{x}_n = (u_n, v_n, \{u_n - v_n\}), \quad n = 1, 2, \ldots,$$

has the a.d.f. $g(\mathbf{x})$ which can be described as follows:
Divide the unit square $[0,1]^2$ into regions $A, B, C, D, E, F, G, H, I$ as shown on the following Figure 2



Then

$$g(x_1,x_2,x_3) = \begin{cases} x_1x_3, & \text{if } (x_1,x_2) \in A, \\ -\frac{1}{2}(x_1^2+x_2^2+x_3^2) + x_1x_2 + x_2x_3, & \text{if } (x_1,x_2) \in B, \\ -\frac{1}{2}x_1^2 + x_1x_2, & \text{if } (x_1,x_2) \in C, \\ \frac{1}{2}x_2^2, & \text{if } (x_1,x_2) \in D, \\ -\frac{1}{2}x_3^2 + x_2x_3, & \text{if } (x_1,x_2) \in E, \\ -\frac{1}{2}x_2^2 + x_1x_2 + x_1x_3 + x_2x_3 - x_1 - x_3 + \frac{1}{2}, & \text{if } (x_1,x_2) \in F, \\ \frac{1}{2}x_1^2 + x_1x_3 + x_2x_3 - x_1 - x_3 + \frac{1}{2}, & \text{if } (x_1,x_2) \in G, \\ \frac{1}{2}(x_1^2+x_2^2+x_3^2) + x_1x_3 - x_1 - x_3 + \frac{1}{2}, & \text{if } (x_1,x_2) \in H, \\ x_1x_2 + x_2x_3 - x_2 & \text{if } (x_1,x_2) \in I. \end{cases}$$

The Weyl criterion implies that the two–dimensional sequence $(u_n, \{u_n - v_n\})$ is u.d., thus the face d.f.s are

$$g(1,x_2,x_3) = x_2x_3, \quad g(x_1,1,x_3) = x_1x_3, \quad g(x_1,x_2,1) = x_1x_2.$$

Another d.f. having these three properties (distinct from the u.d.) is

$$g(x_1, x_2, x_3) = \min(x_1x_2, x_1x_3, x_2x_3).$$

2.10. Ratio sequences

Let $x_1 < x_2 < \dots$ be an increasing sequence of positive integers and consider the sequence of blocks X_n, $n = 1, 2, \dots$, with blocks

$$X_n = \left(\frac{x_1}{x_n}, \frac{x_2}{x_n}, \cdots, \frac{x_n}{x_n}\right)$$

and denote by $F(X_n, x)$ the step d.f.

$$F(X_n, x) = \frac{\#\{i \le n; \frac{x_i}{x_n} < x\}}{n},$$

for $x \in [0, 1)$ and $F(X_n, 1) = 1$. A d.f. g is a d.f. of the sequence of single blocks X_n, if there exists an increasing sequence of positive integers $n_1, n_2, \dots$ such that $\lim_{k\to\infty} F(X_{n_k}, x) = g(x)$ a.e. on $[0, 1]$. Denote by $G(X_n)$ the set of all d.f.s of the sequence of single blocks X_n. $G(X_n)$ has the following properties:

(iii) Assume that all d.f.s in $G(X_n)$ are continuous at 1. Then all d.f.s in $G(X_n)$ are continuous on $(0, 1]$, i.e., only possible discontinuity is in 0.

(iv) If $\underline{d}(x_n) > 0$, then for every $g(x) \in G(X_n)$ we have [45, Th. 6.2(iii)]

$$\frac{\underline{d}(x_n)}{\overline{d}(x_n)}x \le g(x) \le \frac{\overline{d}(x_n)}{\underline{d}(x_n)}x$$

for every $x \in [0, 1]$. Thus $\underline{d}(x_n) = \overline{d}(x_n) > 0$ implies u.d. of the block sequence X_n, $n = 1, 2, \dots$.

(v) If $\underline{d}(x_n) > 0$, then every $g(x) \in G(X_n)$ is continuous on $[0, 1]$.

(vi) If $\underline{d}(x_n) > 0$, then there exists $g(x) \in G(X_n)$ such that $g(x) \ge x$ for every $x \in [0, 1]$, [45, Th. 6.2(ii)]. By [2, Th. 6)] every $G(X_n)$ contains $g(x) \ge x$.

(vii) If $\overline{d}(x_n) > 0$, then there exists $g(x) \in G(X_n)$ such that $g(x) \le x$ for every $x \in [0, 1]$.

(viii) Assume that $G(X_n)$ is singleton, i.e., $G(X_n) = \{g(x)\}$. Then either $g(x) = c_0(x)$ for $x \in [0, 1]$; or $g(x) = x^\lambda$ for some $0 < \lambda \le 1$ and $x \in [0, 1]$. Moreover, if $\overline{d}(x_n) > 0$, then $g(x) = x$.

(ix) $\max_{g\in G(X_n)} \int_0^1 g(x)\mathrm{d}x \ge \frac{1}{2}$.

(x) Assume that every d.f. $g(x) \in G(X_n)$ has a constant value on the fixed interval $(u, v) \subset [0, 1]$ (maybe different). If $\underline{d}(x_n) > 0$ then all d.f.'s in $G(X_n)$ has infinitely many intervals with constant values.

(xi) There exists an increasing sequence x_n, $n = 1, 2, \ldots$, of positive integers such that $G(X_n) = \{h_\alpha(x); \alpha \in [0,1]\}$, where $h_\alpha(x) = \alpha$, $x \in (0,1)$ is the constant d.f.

(xii) There exists an increasing sequence x_n, $n = 1, 2, \ldots$, of positive integers such that $c_1(x) \in G(X_n)$ but $c_0(x) \notin G(X_n)$, where $c_0(x)$ and $c_1(x)$ are one–jump d.f.'s with the jump of height 1 at $x = 0$ and $x = 1$, respectively.

(xiii) There exists an increasing sequence x_n, $n = 1, 2, \ldots$, of positive integers such that $G(X_n)$ is non-connected.

(xiv) $G(X_n) = \{x^\lambda\}$ if and only if $\lim_{n\to\infty}(x_{k.n}/x_n) = k^{1/\lambda}$ for every $k = 1, 2, \ldots$. Here as in (viii) we have $0 < \lambda \leq 1$.

(xv) If $\underline{d}(x_n) > 0$, then all d.f.s $g(x) \in G(X_n)$ are continuous, nonsingular and bounded by $h_1(x) \leq g(x) \leq h_2(x)$, where

$$h_1(x) = \begin{cases} x\frac{\underline{d}}{\overline{d}} & \text{if } x \in \left[0, \frac{1-\overline{d}}{1-\underline{d}}\right], \\ \frac{\underline{d}}{\frac{1}{x}-(1-\underline{d})} & \text{otherwise,} \end{cases} \quad h_2(x) = \min\left(x\frac{\overline{d}}{\underline{d}}, 1\right).$$

Furthermore, there exists x_n, $n = 1, 2, \ldots$, such that $h_2(x) \in G(X_n)$ and for every x_n we have $h_1(x) \notin G(X_n)$, [2, Th. 7] and moreover

(xvi) for a given fixed $g(x) \in G(X_n)$ we have $h_{1,g}(x) \leq g(x) \leq h_{2,g}(x)$, where

$$h_{1,g}(x) = \begin{cases} x\frac{\underline{d}}{d_g} & \text{if } x < y_0 = \frac{1-d_g}{1-\underline{d}}, \\ x\frac{1}{d_g} + 1 - \frac{1}{d_g} & \text{if } y_0 \leq x \leq 1, \end{cases} \tag{78}$$

$$h_{2,g}(x) = \min\left(x\frac{\overline{d}}{d_g}, 1\right), \tag{79}$$

where if $\lim_k \to \infty F(X_{n_k}, x) = g(x)$, then $d_g = \lim_{k\to\infty}\frac{n_k}{x_{n_k}}$ if exists.

Applying (xv) it is proved [2]:

Theorem 32. *For every increasing sequence $x_1 < x_2 < \ldots$ of positive integers with the lower and upper asymptotic densities $0 < \underline{d} \leq \overline{d}$ we have*

$$\frac{1}{2}\frac{\underline{d}}{\overline{d}} \leq \liminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{x_n}, \tag{80}$$

and

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{x_n} \leq \frac{1}{2} + \frac{1}{2}\left(\frac{1-\min(\sqrt{\underline{d}}, \overline{d})}{1-\underline{d}}\right)\left(1 - \frac{\underline{d}}{\min(\sqrt{\underline{d}}, \overline{d})}\right). \tag{81}$$

Here the equations in (80) *and* (81) *can be attained.*

Example 23. O. Strauch and J.T. Tóth (2001): Put $x_n = p_n$, the nth prime and denote

$$X_n = \left(\frac{2}{p_n}, \frac{3}{p_n}, \ldots, \frac{p_{n-1}}{p_n}, \frac{p_n}{p_n} \right).$$

The sequence of blocks X_n is u.d. and therefore the ratio sequence p_m/p_n, $m = 1, 2, \ldots, n$, $n = 1, 2, \ldots$ is u.d. in $[0, 1]$. This generalizes a result of A. Schinzel (cf. W. Sierpiński (1964, p. 155)). Note that from u.d. of X_n applying the L^2 discrepancy of X_n we get the following interesting limit

$$\lim_{n\to\infty} \frac{1}{n^2 p_n} \sum_{i,j=1}^{n} |p_i - p_j| = \frac{1}{3}.$$

The set $G(X_n)$ can be complicated, see F. Filip, L. Mišík and J.T. Tóth [13]:

Example 24. Let a_k, n_k, $k = 1, 2, \ldots$, and x_n, $n = 1, 2, \ldots$ be three increasing integer sequences and $h_1 < h_2$ be two positive integers. Assume that

(i) $\frac{n_k}{n_{k+1}} \to 0$ for $k \to \infty$;

(ii) $\frac{a_k}{n_{k+1}} \to 0$ for $k \to \infty$;

(iii) for odd k we have
$a_k^{h_2} \le x_{n_k} = (a_{k-1} + n_k - n_{k-1})^{h_1} \le (a_k + 1)^{h_2}$ and
$x_i = (a_k + i - n_k)^{h_2}$ for $n_k < i \le n_{k+1}$;

(iv) for even k we have
$a_k^{h_1} \le x_{n_k} = (a_{k-1} + n_k - n_{k-1})^{h_2} \le (a_k + 1)^{h_1}$ and
$x_i = (a_k + i - n_k)^{h_1}$ for $n_k < i \le n_{k+1}$.

Then $\frac{x_n}{x_{n+1}} \to 1$ and the set $G(X_n)$ of all distribution functions of the sequence of blocks X_n is $G(X_n) = G_1 \cup G_2 \cup G_3 \cup G_4$, where

$G_1 = \{x^{\frac{1}{h_2}} \cdot t; t \in [0, 1]\}$,

$G_2 = \{x^{\frac{1}{h_2}}(1 - t) + t; t \in [0, 1]\}$,

$G_3 = \{\max(0, x^{\frac{1}{h_1}} - (1 - x^{\frac{1}{h_1}})u); u \in [0, \infty)\}$ and

$G_4 = \{\min(1, x^{\frac{1}{h_1}} \cdot v); v \in [1, \infty)\}$.

In Unsolved Problems [42, 1.9] there are given many open questions on $G(X_n)$, one from them: Characterize a nonempty set H of d.f.s for which there exists an increasing sequence of positive integers x_n such that $G(X_n) = H$. In [2] is given the following partial result:

Theorem 33. *Let H be a nonempty set of d.f.s defined on $[0, 1]$. Then there exists a positive integer sequence $x_1 < x_2 < \ldots$ such that $H \subset G(X_n)$.*

3. Calculation methods of $G(x_n)$

3.1. Calculation of d.f.s by definition

We give a proof of Example 6 via definition of d.f.s.

Example 25. Starting with $x_n = \{\log\log n\}$ all the sequences $\{\log\log\ldots\log n\}$ have

$$G(x_n) = \{c_\alpha(x); \alpha \in [0,1]\} \cup \{h_\alpha(x); \alpha \in [0,1]\}.$$

P r o o f. For the first iterated logarithm we chose an index-sequence N_k as

$$N_k = [\exp\exp(k+\alpha)]$$

Then we have $\lim\limits_{k\to\infty} F_{N_k}(x) = c_\alpha(x)$. For $N_k = [\exp\exp(k+\varepsilon_k)]$, where $\varepsilon_k \to 0$ such that $\big(\exp\exp(k+\varepsilon_k)\big)/(\exp\exp k) \to \beta$, we have $\lim\limits_{k\to\infty} F_{N_k}(x) = h_\alpha(x)$, where $\alpha = (\beta-1)/\beta$.

On the other hand, let $\lim\limits_{n\to\infty} F_{N_n}(x) = g(x)$. Then $N_n = \exp\exp(k_n+\varepsilon_n)$, where $k_n = [\log\log N_n]$, $\varepsilon_n = \{\log\log N_n\}$, and the sequence $(\varepsilon_n)_{n=1}^\infty$ cannot have different limit points. □

Similarly, but complicated, it can be found:

Example 26. The set of all d.f.s of the two-dimensional sequence

$$(\log n, \log\log n) \bmod 1$$

has the form, see [43]:

$$\begin{aligned} G((\log n, \log\log n) \bmod 1) = {} & \{g_{u,v}(x,y); u \in [0,1], v \in [0,1]\} \\ & \cup \{g_{u,0,j,\alpha}(x,y); u \in [0,1], \alpha \in A, j = 1,2,\ldots\} \\ & \cup \{g_{u,0,0,\alpha}(x,y); u \in [\alpha,1], \alpha \in A\}, \end{aligned} \tag{82}$$

where A is the set of all limit points of the sequence $e^n \bmod 1$, $n = 1,2,\ldots$, and, [15] for $(x,y) \in [0,1)^2$,

$$\begin{aligned} g_{u,v}(x,y) &= g_u(x) \cdot c_v(y), \\ g_{u,0,j,\alpha}(x,y) &= g_{u,0,j,\alpha}(x) \cdot c_0(y), \\ g_{u,0,0,\alpha}(x,y) &= g_{u,0,0,\alpha}(x) \cdot c_0(y), \end{aligned}$$

where

$$g_u(x) = \frac{e^{\min(x,u)}-1}{e^u} + \frac{1}{e^u}\frac{e^x-1}{e-1},$$

[15] The exact form of A is a well-known open problem.

$$c_v(y) = \begin{cases} 0 & \text{if } 0 \le y \le v, \\ 1 & \text{if } v < y \le 1, \end{cases} \qquad \text{and} \qquad c_v(0) = 0, c_v(1) = 1,$$

$$g_{u,0,j,\alpha}(x) = \frac{e^{\max(\alpha,x)} - e^{\alpha}}{e^{j+u}} + \frac{e^{\min(x,u)} - 1}{e^u} + \frac{1}{e^u}\frac{e^x - 1}{e - 1}\left(1 - \frac{1}{e^{j-1}}\right),$$

$$g_{u,0,0,\alpha}(x) = \frac{e^{\max(\min(x,u),\alpha)} - e^{\alpha}}{e^u}. \tag{83}$$

A proof via d.f.s. consider limits of

$$F_N(x,y) = \frac{\#\{3 \le n \le N; (\{\log n\}, \{\log\log n\}) \in [0,x) \times [0,y)\}}{N}$$
$$= \frac{\sum_{j=0}^{J} \#\{[e^{e^j}, e^{e^{j+y}}) \cap (\cup_{k=1}^{K} [e^k, e^{k+x})) \cap \mathbf{N}\}}{N},$$

where

$$K = [\log N], \qquad J = [\log\log N], \qquad \mathbf{N} = \{1, 2, \ldots, N\}.$$

O. Strauch and O. Blažeková in [43] and R. Giuliano Antonini and O. Strauch in [15] generalized Example 26 and Koksma [19], [20, Kap. 8] (cf. [22, p. 58, Th. 7.7]) to the following Theorems 34, 35 and 36. Assume:

(I) $f(x)$ be a real-valued function defined for $x \ge 1$ such that $f(x)$ is strictly increasing with inverse function $f^{-1}(x)$.

(II) $\lim_{k\to\infty} \frac{f^{-1}(k+x) - f^{-1}(k)}{f^{-1}(k+1) - f^{-1}(k)} = \tilde{g}(x)$ for each $x \in [0,1]$, point of continuity of $\tilde{g}(x)$;

(III) $\lim_{k\to\infty} \frac{f^{-1}(k+u)}{f^{-1}(k)} = \psi(u)$ for each $u \in [0,1]$, point of continuity of $\psi(u)$, or $\psi(u) = \infty$ for $u > 0$;

(IV) $\lim_{k\to\infty} f^{-1}(k+1) - f^{-1}(k) = \infty$.

For computing $G(f(n) \bmod 1)$ we have the following three theorems.

Theorem 34. *If $1 < \psi(1) < \infty$ and $f'(x) \to 0$ as $x \to \infty$, then*

$$G(f(n) \bmod 1) = \left\{ g_u(x) = \frac{\min(\psi(x), \psi(u)) - 1}{\psi(u)} + \frac{1}{\psi(u)}\tilde{g}(x); u \in [0,1] \right\}, \tag{84}$$

where

$$\tilde{g}(x) = \frac{\psi(x) - 1}{\psi(1) - 1} \quad \text{and} \quad F_{N_i}(x) \to g_u(x) \quad \text{as} \quad i \to \infty$$

if and only if $f(N_i) \bmod 1 \to u$.

The lower d.f. $\underline{g}(x)$ *and the upper d.f.* $\overline{g}(x)$ *of* $f(n) \bmod 1$ *are*

$$\underline{g}(x) = \tilde{g}(x), \quad \overline{g}(x) = 1 - \frac{1}{\psi(x)}\big(1 - \tilde{g}(x)\big).$$

Furthermore $\underline{g}(x) = g_0(x) = g_1(x)$ *belongs to* $G(f(n) \bmod 1)$ *but* $\overline{g}(x) = g_x(x)$ *does not.*

Theorem 35. *If* $\psi(1) = 1$, *then the sequence* $f(n) \bmod 1$, $n = 1, 2, \ldots$ *has a.d.f.* $\tilde{g}(x)$, *i.e.,*

$$G(f(n) \bmod 1) = \{\tilde{g}(x)\}. \tag{85}$$

Theorem 36. *Let* $\psi(u) = \infty$, *for every* $u > 0$ *and for* $u = 0$ *the limit* $\psi(u)$ *is not defined in the way that for every* $t \in [0, \infty)$ *there exists a sequence* $u(k) \to 0$ *such that* (i) $\lim_{k\to\infty} \frac{f^{-1}(k+u(k))}{f^{-1}(k)} = t$. *Then we have*

$$G(f(n) \bmod 1) = \{c_u(x); u \in [0,1]\} \cup \{h_\beta(x); \beta \in [0,1]\}, \tag{86}$$

where $F_{N_i} \to c_u(x)$ *if and only if* $f(N_i) \bmod 1 \to u > 0$ *and* $F_{N_i} \to h_\beta(x)$ *if and only if* $f(N_i) \bmod 1 \to 0$ *and* $\frac{f^{-1}([f(N_i)])}{N_i} \to 1 - \beta$.

In proofs of Theorems 34, 35 and 36 there are studied limits of step d.f. $F_N(x)$ expressed as

$$F_N(x) = \frac{\sum_{k=0}^{K-1} A_N\big([k, k+x)\big)}{N} + \frac{A_N\big([K, K+x) \cap [K, K+w)\big)}{N} + \frac{O\left(A_N\big([1, f^{-1}(0))\big)\right)}{N},$$

where

$$K = [f(N)], \qquad w = \{f(N)\}, \qquad A_N\big([x,y)\big) = \#\{n \le N; f(n) \in [x,y)\}.$$

Example 27. Applying Theorem 34 to the function $f(x) = \log(x \log^{(i)} x)$, where $\log^{(i)} x$ is the ith iterated logarithm $\log \ldots \log x$, we find

$$G\big(\log(n \log^{(i)} n) \bmod 1\big) = G(\log n \bmod 1).$$

3.2. Connectivity of $G(x_n)$

The connectivity of $G(x_n)$ implies the following simple theorem: Define

• For the d.f. g the Graph(g) be the continuous curve formed by all the points $(x, g(x))$ for $x \in [0,1]$, and the all line segments connecting the points of discontinuity $(x, \liminf_{x'\to x} g(x'))$ and $(x, \limsup_{x'\to x} g(x'))$.

• Denote $\underline{g}_H(x) = \inf_{g\in H} g(x)$ and $\overline{g}_H(x) = \sup_{g\in H} g(x)$.

Theorem 37 ([37]). *Let H be a non–empty, closed, and connected set of d.f.s. Assume that for every $g \in H$ there exists a point $(x,y) \in \mathrm{Graph(g)}$ such that $(x,y) \notin \mathrm{Graph}(\widetilde{\mathrm{g}})$ for any $\widetilde{g} \in H$ with $\widetilde{g} \neq g$. If*

(i) $\underline{g} = \underline{g}_H$ *and* $\overline{g} = \overline{g}_H$ *for the lower d.f.* $\underline{g}$ *and the upper d.f.* $\overline{g}$ *of the sequence* $x_n \in [0,1)$, *and*

(ii) $G(x_n) \subset H$,

then $G(x_n) = H$.

To prove $G(x_n) \subset H$ it can be used:

Theorem 38. *Let $F(x,y)$ be a real continuous function defined on $[0,1]^2$ and $G(F)$ is the set of all d.f. $g(x)$ which satisfy $\int_0^1 \int_0^1 F(x,y)\mathrm{d}g(x)\mathrm{d}g(y) = 0$. Then for every sequence $x_n \in [0,1)$ we have*

$$G(x_n) \subset G(F) \Longleftrightarrow \lim_{N\to\infty} \frac{1}{N^2} \sum_{m,n=1}^{N} F(x_m, x_n) = 0. \tag{87}$$

P r o o f. Using the definition of the Riemann-Stieltjes integral, we have

$$\int_0^1 \int_0^1 F(x,y)\mathrm{d}F_N(x)\mathrm{d}F_N(y) = \frac{1}{N^2} \sum_{m,n=1}^{N} F(x_m, x_n).$$

Suppose that $\lim_{k\to\infty} F_{N_k}(x) = g(x)$ for all continuity points x of g. Then, applying the Helly-Bray lemma, we find

$$\int_0^1 \int_0^1 F(x,y)\mathrm{d}F_{N_k}(x)\mathrm{d}F_{N_k}(y) = \int_0^1 \int_0^1 F(x,y)\mathrm{d}g(x)\mathrm{d}g(y),$$

and the implication $\Longleftarrow$ in (87) follows immediately.

In order to show the implication $\Longrightarrow$, assume

$$\lim_{k\to\infty} \frac{1}{{N_k}^2} \sum_{m,n=1}^{N_k} F(x_m, x_n) = \beta > 0.$$

By the Helly selection principle there exists a subsequence N_k' of N_k such that

$$\lim_{k\to\infty} F_{N_k'}(x) = g(x) \in G(x_n).$$

Again, by the Helly-Bray theorem we find $\int_0^1 \int_0^1 F(x,y)\mathrm{d}g(x)\mathrm{d}g(y) = \beta$. We conclude $g \notin G(F)$. □

Theorem 37 implies:

• Let $g_1 \neq g_2$ be two d.f.s. Denote

$$F_{g_2}(x,y) = \int_0^x g_2(t)\mathrm{d}t + \int_0^y g_2(t)\mathrm{d}t - \max(x,y) + \int_0^1 (1-g_2(t))^2\mathrm{d}t,$$

$$F_{g_1,g_2}(x) = \frac{\int_0^x \big(g_2(t)-g_1(t)\big)\mathrm{d}t - \int_0^1 \big(1-g_2(t)\big)\big(g_2(t)-g_1(t)\big)\mathrm{d}t}{\int_0^1 \big(g_2(t)-g_1(t)\big)^2\mathrm{d}t},$$

$$F_{g_1,g_2}(x,y) = F_{g_2}(x,y) - F_{g_1,g_2}(x)F_{g_1,g_2}(y).\int_0^1 \big(g_2(t)-g_1(t)\big)^2\mathrm{d}t.$$

Theorem 39. *For given sequence $x_n \in [0,1)$ we have*

$$G(x_n) = \{tg_1(x) + (1-t)g_2(x); t \in [0,1]\}$$

if and only if

(i) $\lim_{N\to\infty} \frac{1}{N^2} \sum_{m,n=1}^{N} F_{g_1,g_2}(x_m,x_n) = 0$,

(ii) $\liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} F_{g_1,g_2}(x_n) = 0$,

(iii) $\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} F_{g_1,g_2}(x_n) = 1$.

Example 28. Put

$$F_1(x,y) = 1 - \max(x,y) - \frac{3}{4}(1-x^2)(1-y^2),$$

$$F_2(x,y) = \frac{x+y}{2} - \max(x,y) + \frac{1}{4} - 3(x-x^2)(y-y^2),$$

$$F_3(x,y) = 1 - \max(x,y),$$

$$F_4(x,y) = \frac{x+y}{2} - \max(x,y) + \frac{1}{4},$$

and

$$H_1 = \{tx + (1-t)c_1(x); t \in [0,1]\}, \quad \text{and} \quad H_2 = \{tx + (1-t)h_{1/2}(x); t \in [0,1]\}.$$

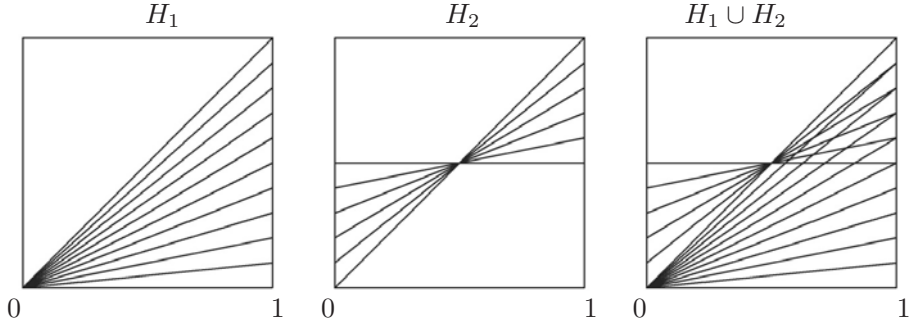


Figure 1: D.f.s H_1, H_2 and $H_1 \cup H_2$.

Then $G(x_n) = H_1 \cup H_2$ for a sequence x_n in $[0,1)$ if and only if

(i) $\lim_{N\to\infty} \frac{1}{N^4} \sum_{m,n,k,l=1}^{N} F_1(x_m, x_n)F_2(x_k, x_l) = 0$,

(ii) $\liminf_{N\to\infty} \frac{1}{N^2} \sum_{m,n=1}^{N} F_3(x_m, x_n) = 0$.

(iii) $\liminf_{N\to\infty} \frac{1}{N^2} \sum_{m,n=1}^{N} F_4(x_m, x_n) = 0$.

Here $c_1(x)$ is the one–jump d.f. with jump at $x = 1$ and $h_{1/2}(x)$ is the d.f. taking constant value $1/2$.

3.2.1. The moment problem $\int_0^1 \int_0^1 F(x,y)\mathrm{d}g(x)\mathrm{d}g(y) = 0$

These investigation is motivated by Theorem 37. Again, for a given continuous $F : [0,1]^2 \to \mathbb{R}$, let $G(F)$ denote the set of all distribution functions g which solve the following moment problem $\int_0^1 \int_0^1 F(x,y)\mathrm{d}g(x)\mathrm{d}g(y) = 0$.

Theorem 40 ([37]). *Let $F : [0,1]^2 \to \mathbb{R}$ be a continuous and symmetric function. For every distribution functions $g(x)$, $\tilde{g}(x)$ we have*

$$\int_0^1 \int_0^1 F(x,y)\mathrm{d}g(x)\mathrm{d}g(y) = 0 \Longleftrightarrow \int_0^1 \int_0^1 F(x,y)\mathrm{d}\tilde{g}(x)\mathrm{d}\tilde{g}(y)$$
$$= \int_0^1 \big(g(x) - \tilde{g}(x)\big)\Big(2\mathrm{d}_x F(x,1) - \int_0^1 \big(g(y) + \tilde{g}(y)\big)\mathrm{d}_y\mathrm{d}_x F(x,y)\Big) \quad (88)$$

Example 29. Putting $\tilde{g}(x) = c_0(x)$ in (88), we have

$$\int_0^1 \int_0^1 F(x,y)\mathrm{d}g(x)\mathrm{d}g(y) = 0 \Longleftrightarrow$$
$$F(0,0) = \int_0^1 \big(g(x) - 1\big)\Big(2\mathrm{d}_x F(x,1) - \int_0^1 \big(g(y) + 1\big)\mathrm{d}_y\mathrm{d}_x F(x,y)\Big). \quad (89)$$

For $F(x,y) = F_0(x,y)$ we have

$$g(x) = x \iff \frac{1}{3} = \int_0^1 g(x)\big(2x - g(x)\big)\mathrm{d}x.$$

Here

$$F_0(x,y) = \frac{1}{3} + \frac{x^2+y^2}{2} - \max(x,y)$$
$$= \frac{1}{3} + \frac{x^2+y^2}{2} - \frac{x+y}{2} - \frac{|x-y|}{2}$$

This function $F_0(x,y)$ is inspired by the classical L^2 discrepancy

$$\int_0^1 \big(F_N(x)-x\big)^2 \mathrm{d}x$$

since

$$\int_0^1 \big(F_N(x)-x\big)^2 \mathrm{d}x = \frac{1}{N^2}\sum_{m,n=1}^{N} F_0(x_m,x_n).$$

3.3. Computation $G\big(h(x_n,y_n)\big)$ by $g(x,y)\in G\big((x_n,y_n)\big)$

We describe a result for $h(x,y)=x+y \bmod 1$.

Theorem 41. *Let x_n and y_n be two sequences in $[0,1)$ and $G((x_n,y_n))$ denote the set of all d.f.s of the two-dimensional sequence (x_n,y_n). If*

$$z_n = x_n + y_n \bmod 1,$$

then the set $G(z_n)$ of all d.f.s of z_n has the form

$$G(z_n) = \left\{g(t)=\int_{0\le x+y<t} 1.\mathrm{d}g(x,y)+\int_{1\le x+y<1+t} 1.\mathrm{d}g(x,y);\ \ g(x,y)\in G\big((x_n,y_n)\big)\right\}$$

assuming that all the used Riemann-Stieltjes integrals exist.

Example 30. Applying Theorem 41 to the $G\big((\log n,\log\log n) \bmod 1\big)$ in Example 26 it can be found, for

$$G\big(\log(n\log n) \bmod 1\big) = \{g_{u,v}(x); u\in[0,1], v\in[0,1]\},$$

that

$$g_{u,v}(x) = \begin{cases} g_u(1+x-v)-g_u(1-v) & \text{if } 0\le x\le v,\\ g_u(x-v)+1-g_u(1-v) & \text{if } v<x\le 1.\end{cases}$$

Directly by means of computation we see that

$$g_{u,v}(x) = g_w(x), \quad \text{for} \quad w = u+v \bmod 1,$$

defined in Example 1 and thus $G\big(\log(n\log n) \bmod 1\big) = G\big(\log n \bmod 1\big)$. The same holds by Example 27.

3.4. Solution of $(X_1,X_2,X_3) = \Big(\int_0^1 g(x)\mathrm{d}x, \int_0^1 xg(x)\mathrm{d}x, \int_0^1 g^2(x)\mathrm{d}x\Big)$

See [39], [44, 2.2.21]:

It is motivated by L^2 discrepancy criterion for g-distributed sequences: A sequence x_n in $[0,1]$ has a.d.f. $g(x)$ if and only if

$$\lim_{N\to\infty}\left(1+\int_0^1 g^2(x)\mathrm{d}x-2\int_0^1 g(x)\mathrm{d}x+\frac{2}{N}\sum_{n=1}^{N}\int_0^{x_n} g(x)\mathrm{d}x - \frac{1}{N}\sum_{n=1}^{N} x_n-\frac{1}{2N^2}\sum_{m,n=1}^{N}|x_m-x_n|\right)=0,$$

or equivalently, if and only if

(i) $\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} x_n=\int_0^1 x\mathrm{d}g(x)$,

(ii) $\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\int_0^{x_n} g(x)\mathrm{d}x=\int_0^1\left(\int_0^x g(t)\mathrm{d}t\right)\mathrm{d}g(x)$,

(iii) $\lim_{N\to\infty}\frac{1}{N^2}\sum_{m,n=1}^{N}|x_m-x_n|=\int_0^1\int_0^1|x-y|\mathrm{d}g(x)\mathrm{d}g(y)$.

Since the left-hand side of (iii) contains $g(x)$ we shall instead it by the second moment and we solve

$$(s_1,s_2,s_3)=\left(\int_0^1 x\mathrm{d}g(x),\int_0^1 x^2\mathrm{d}g(x),\int_0^1\int_0^1|x-y|\mathrm{d}g(x)\mathrm{d}g(y)\right), \tag{90}$$

where

$s_1=\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} x_n$,

$s_2=\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} x_n^2$ and

$s_3=\lim_{N\to\infty}\frac{1}{N^2}\sum_{m,n=1}^{N}|x_m-x_n|$.

Using $(s_1,s_2,s_3)=(1-X_1,1-2X_2,2(X_1-X_3))$ then (90) can be transform to

$$(X_1,X_2,X_3)=\left(\int_0^1 g(x)\mathrm{d}x,\int_0^1 xg(x)\mathrm{d}x,\int_0^1 g^2(x)\mathrm{d}x\right).$$

Define, for every nondecreasing $g:[0,1]\to[0,1]$, the operator

$$\mathbf{F}(g)=\left(\int_0^1 g(x)\mathrm{d}x,\int_0^1 xg(x)\mathrm{d}x,\int_0^1 g^2(x)\mathrm{d}x\right).$$

For $\mathbf{F}$, we introduce the body

$$\mathbf{\Omega}=\big\{\mathbf{F}(g);g\,[0,1]\to[0,1],g\text{ is nondecreasing}\big\},$$

and $\partial\mathbf{\Omega}$ denote the *boundary* of $\mathbf{\Omega}$.

Let $g(u_1,v_1,u_2,v_2)$ denote the distribution function $h(x)$ defined by

$$h(x)=\begin{cases}0 & \text{for } 0\le x\le v_1,\\ \frac{u_2-u_1}{v_2-v_1}x+u_1-v_1\frac{u_2-u_1}{v_2-v_1} & \text{for } v_1<x\le v_2,\\ 1 & \text{for } v_2<x\le 1\end{cases}$$

and put $\mathbf{X}=(X_1,X_2,X_3)$. O. Strauch [39] proved

Theorem 42. *For the moment problem* $\mathbf{X} = \mathbf{F}(g)$*, to have only a finite number of solutions in distribution functions* g *it is necessary and sufficient that* $\mathbf{X} \in \partial\mathbf{\Omega}$*. We express the boundary* $\partial\mathbf{\Omega}$ *as*

$$\partial\mathbf{\Omega} = \bigcup_{1\le i\le 7} \Pi_i.$$

In addition, for $\mathbf{X} \in \Pi_i$*,* $i = 1, 2, \dots, 6$*, the moment problem* $\mathbf{X} = \mathbf{F}(g)$ *is uniquely solvable as* $g = g^{(i)}$*, and for* $\mathbf{X} \in \Pi_7$ *has precisely two solutions of types* $g^{(7)}$ *and* $g^{(7^*)}$*.*

Theorem 43. *Let* x_n*,* $n = 1, 2, \dots$ *be a sequence in* $[0, 1]$ *with the limits*

$X_1 = 1 - \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} x_n,$

$X_2 = \frac{1}{2} - \frac{1}{2} \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} x_n^2,$

$X_3 = 1 - \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} x_n - \frac{1}{2} \lim_{N\to\infty} \frac{1}{N^2} \sum_{m,n=1}^{N} |x_m - x_n|.$

If

$$\mathbf{X} = (X_1, X_2, X_3) \in \bigcup_{1\le i\le 7} \Pi_i,$$

then the sequence x_n *has an a.d.f. Precisely, if*

$$\mathbf{X} \in \Pi_i, \quad i = 1, \dots, 6,$$

then x_n *has a.d.f.* $g^{(i)}$*, and if*

$$\mathbf{X} \in \Pi_7,$$

then x_n *has a.d.f. either* $g^{(7)}$ *or* $g^{(7^*)}$*, depending on whether*

$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \int_0^{x_n} g^{(7)}(t)\mathrm{d}t = X_1 - X_3$ *or*

$\lim_{n\to\infty} \frac{1}{N} \sum_0^N \int_0^{x_n} g^{(7^*)}(t)\mathrm{d}t = X_1 - X_3.$

Here

$$\Pi_1 = \Big\{(X_1, X_2, X_3); \quad X_2 = \frac{1}{2} - \frac{1}{2}(1 - X_1)^2 - \frac{3}{2}(X_1 - X_3)^2, \\ \max\Big(\frac{4}{3}X_1 - \frac{1}{3},\ \frac{2}{3}X_1\Big) \le X_3 \le X_1,\ 0 \le X_1 \le 1\Big\},$$

$$\Pi_2 = \Big\{(X_1, X_2, X_3); \quad X_2 = \frac{1}{2}X_1 + \frac{1}{2}\sqrt{\frac{1}{3}(X_3 - X_1^2)}, \\ X_1^2 \le X_3 \le \min\Big(\frac{4}{3}X_1^2,\ \frac{4}{3}X_1^2 - \frac{2}{3}X_1 + \frac{1}{3}\Big),\ 0 \le X_1 \le 1\Big\},$$

$$\Pi_3 = \Big\{(X_1,X_2,X_3);\quad X_2 = \frac{1}{2} - \frac{4}{9}\frac{(1-X_1)^3}{(1+X_3-2X_1)},$$
$$\frac{4}{3}X_1^2 - \frac{2}{3}X_1 + \frac{1}{3} \le X_3 \le \frac{4}{3}X_1 - \frac{1}{3},\ \frac{1}{2} \le X_1 \le 1\Big\},$$

$$\Pi_4 = \Big\{(X_1,X_2,X_3);\quad X_2 = X_1 - \frac{4}{9}\frac{X_1^3}{X_3}, \frac{4}{3}X_1^2 \le X_3 \le \frac{2}{3}X_1,\ 0 \le X_1 \le \frac{1}{2}\Big\},$$

$$\Pi_5 = \Big\{(X_1,X_2,X_3);\quad X_2 = \frac{1}{2} - \frac{1}{2}\frac{(1-X_1)^3}{(1+X_3-2X_1)},$$
$$X_1^2 \le X_3 \le X_1,\ 0 \le X_1 < \frac{1}{2}\Big\},$$

$$\Pi_6 = \Big\{(X_1,X_2,X_3);\quad X_2 = X_1 - \frac{1}{2}\frac{X_1^3}{X_3},\ X_1^2 \le X_3 \le X_1,\ \frac{1}{2} < X_1 \le 1\Big\},$$

$$\Pi_7 = \Big\{\Big(\frac{1}{2}, \frac{1}{2} - \frac{1}{16X_3}, X_3\Big);\quad \frac{1}{4} < X_3 < \frac{1}{2}\Big\},$$

and

$$g^{(1)} = g(0, (1-X_1) - 3(X_1 - X_3), 1, (1-X_1) + 3(X_1 - X_3)),$$
$$g^{(2)} = g\left(X_1 - \sqrt{3(X_3 - X_1^2)}, 0, X_1 + \sqrt{3(X_3 - X_1^2)}, 1\right),$$
$$g^{(3)} = g\left(1 - \frac{3}{2}\frac{1+X_3-2X_1}{1-X_1}, 0, 1, \frac{4}{3}\frac{(1-X_1)^2}{(1+X_3-2X_1)}\right),$$
$$g^{(4)} = g\left(0, 1 - \frac{4X_1^2}{3X_3}, \frac{3X_3}{2X_1}, 1\right),$$
$$g^{(5)} = g\left(\frac{X_1 - X_3}{1-X_1}, 0, \frac{X_1 - X_3}{1-X_1}, \frac{(1-X_1)^2}{1+X_3-2X_1}\right),$$
$$g^{(6)} = g\left(\frac{X_3}{X_1}, 1 - \frac{X_1^2}{X_3}, \frac{X_3}{X_1}, 1\right),$$
$$g^{(7)} = g\left(1 - 2X_3, 0, 1 - 2X_3, \frac{1}{4X_3}\right),$$
$$g^{(7^*)} = g\left(2X_3, 1 - \frac{1}{4X_3}, 2X_3, 1\right).$$

Example 31. Since the straight line $X_2 - X_3 = \frac{1}{12}$ is touched to the projection of $\mathbf{\Omega}$ in $X_2 \times X_3$, then

$$\max_{g(x)-\text{ d.f.}} \int_0^1 \big(x - g(x)\big)g(x)\mathrm{d}x = \frac{1}{12}$$

and it is attained in $g(x) = \frac{x}{2}$.

Example 32. The points[16]

(X_1, X_2, X_3) for $g(x) = c_\alpha(x)$ and (X_1, X_2, X_3) for $g(x) = h_\alpha(x)$, $\alpha \in [0,1]$, lie on the contour of $\mathbf{\Omega}$ and for such (X_1, X_2, X_3) the d.f. $g(x)$ is given uniquely. This implies

$$\int_0^1 x^2\mathrm{d}g(x) = \left(\int_0^1 x\mathrm{d}g(x)\right)^2 \Longleftrightarrow \exists_{\alpha\in[0,1]} g(x) = c_\alpha(x),$$

$$\int_0^1 x^2\mathrm{d}g(x) = \int_0^1 x\mathrm{d}g(x) \Longleftrightarrow \exists_{\alpha\in[0,1]} g(x) = h_\alpha(x),$$

$$\left(1 - \int_0^1 x\mathrm{d}g(x)\right)\left(\int_0^1 x\mathrm{d}g(x)\right) = \frac{1}{2}\int_0^1\int_0^1 |x-y|\mathrm{d}g(x)\mathrm{d}g(y) \Longleftrightarrow \exists_{\alpha\in[0,1]} g(x) = h_\alpha(x)$$

and we also have

$$G(x_n) \subset \{c_\alpha(x); \alpha \in [0,1]\} \Longleftrightarrow \lim_{N\to\infty}\left(\left(\frac{1}{N}\sum_{n=1}^N x_n\right)^2 - \frac{1}{N}\sum_{n=1}^N x_n^2\right) = 0,$$

$$G(x_n) \subset \{h_\alpha(x); \alpha \in [0,1]\} \Longleftrightarrow \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^N (x_n - x_n^2) = 0,$$

$$G(x_n) \subset \{h_\alpha(x); \alpha \in [0,1]\} \Longleftrightarrow \lim_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^N x_n - \left(\frac{1}{N}\sum_{n=1}^N x_n\right)^2 - \frac{1}{N^2}\sum_{m,n=1}^N |x_m - x_n|\right) = 0.$$

3.5. Mapping x_n to $f(x_n)$

Let $f : [0,1] \to [0,1]$ be a function such that, for all $x \in [0,1]$, $f^{-1}([0,x))$ can be expressed as a sum of finitely many pairwise disjoint subintervals $I_i(x)$ of $[0,1]$ with endpoints $\alpha_i(x) \le \beta_i(x)$. For any d.f. $g(x)$ we put

$$g_f(x) = \sum_i g\big(\beta_i(x)\big) - g\big(\alpha_i(x)\big) = \int_{f^{-1}\left([0,x)\right)} 1.\mathrm{d}g(x).$$

[16] $h_\alpha(x) = \alpha$ for $x \in (0,1)$.

The mapping $g \to g_f$ can be used for the studying $G(x_n)$ by the following statement:

Theorem 44. *Let $x_n \bmod 1$ be a sequence having $g(x)$ as a d.f. associated with the sequence of indices $N_1, N_2, \dots$ Suppose that any term $x_n \bmod 1$ is repeated only finitely many times. Then the sequence $f(\{x_n\})$ has the d.f.s $g_f(x)$ for the same $N_1, N_2, \dots$, and vice-versa any distribution function of $f(\{x_n\})$ has this form.*

For example, Theorem 44 can be used to study the sequence $\xi(3/2)^n \bmod 1$, $n = 1, 2, \dots$ Consider

$$f(x) = 2x \bmod 1, \text{ and } h(x) = 3x \bmod 1.$$

In this case, for every $x \in [0, 1]$, we have

$$g_f(x) = g\big(f_1^{-1}(x)\big) + g\big(f_2^{-1}(x)\big) - g(1/2),$$

$$g_h(x) = g\big(h_1^{-1}(x)\big) + g\big(h_2^{-1}(x)\big) + g\big(h_3^{-1}(x)\big) - g(1/3) - g(2/3),$$

with inverse functions

$$f_1^{-1}(x) = x/2, \quad f_2^{-1}(x) = (x+1)/2,$$

and

$$h_1^{-1}(x) = x/3, \quad h_2^{-1}(x) = (x+1)/3, \quad h_3^{-1}(x) = (x+2)/3.$$

Pjateckiĭ-Šapiro [31], by means of the ergodic theory, proved that a necessary and sufficient condition that the sequence $\xi q^n \bmod 1$ with integer$q > 1$ has a distribution function $g(x)$ is that $g_\varphi(x) = g(x)$ for all $x \in [0, 1]$, where $\varphi(x) = qx \bmod 1$. For $\xi(3/2)^n \bmod 1$ we have the following similar necessity.

Theorem 45. *Any distribution function $g(x)$ of $\xi(3/2)^n \bmod 1$ satisfies the functional equation $g_f(x) = g_h(x)$ for all $x \in [0, 1]$.*

The above theorem yields to the following sets of uniqueness for distribution functions of $\xi(3/2)^n \bmod 1$.

Theorem 46. *Let g_1, g_2 be any two distribution functions satisfying $g_{i_f}(x) = g_{i_h}(x)$ for $i = 1, 2$ and $x \in [0, 1]$. Denote*

$$I_1 = [0, 1/3], \quad I_2 = [1/3, 2/3], \quad I_3 = [2/3, 1].$$

If $g_1(x) = g_2(x)$ for $x \in I_i \cup I_j$, $1 \le i \ne j \le 3$, then $g_1(x) = g_2(x)$ for all $x \in [0, 1]$.

Next we have an integral formula for testing $g_f = g_h$. Denote

$$F(x, y) = |\{2x\} - \{3y\}| + |\{2y\} - \{3x\}| - |\{2x\} - \{2y\}| - |\{3x\} - \{3y\}|.$$

Theorem 47. *The continuous distribution function g satisfies $g_f = g_h$ on $[0,1]$ if and only if $\int_0^1 \int_0^1 F(x,y)\mathrm{d}g(x)\mathrm{d}g(y) = 0$.*

The following theorem (see O. Strauch [40]) can be used for generating solutions of $g_f = g_h$.

Theorem 48. *Let g_1, g_2 be two absolutely continuous distribution functions satisfying $g_{1_h}(x) = g_{2_f}(x)$ for $x \in [0,1]$. Then the absolutely continuous distribution function $g(x)$ satisfies $g_f(x) = g_1(x)$ and $g_h(x) = g_2(x)$ for $x \in [0,1]$ if and only if $g(x)$ has the form*

$$g(x) = \begin{cases} \Psi(x), & \text{for } x \in [0,1/6], \\ \Psi(1/6) + \Phi(x-1/6), & \text{for } x \in [1/6,2/6], \\ \Psi(1/6) + \Phi(1/6) + g_1(1/3) - \Psi(x-2/6) & \\ +\Phi(x-2/6) - g_1(2x-1/3) + g_2(3x-1), & \text{for } x \in [2/6,3/6], \\ 2\Phi(1/6) + g_1(1/3) - g_1(2/3) + g_2(1/2) & \\ -\Psi(x-3/6) + g_1(2x-1), & \text{for } x \in [3/6,4/6], \\ -\Psi(1/6) + 2\Phi(1/6) + g_1(1/3) - g_1(2/3) + g_2(1/2) & \\ -\Phi(x-4/6) + g_1(2x-1), & \text{for } x \in [4/6,5/6], \\ -\Psi(1/6) + \Phi(1/6) + g_1(1/3) + \Psi(x-5/6) & \\ -\Phi(x-5/6) - g_1(2x-5/3) + g_2(3x-2), & \text{for } x \in [5/6,1], \end{cases}$$

where

$$\Psi(x) = \int_0^x \psi(t)\mathrm{d}t, \quad \Phi(x) = \int_0^x \phi(t)\mathrm{d}t, \text{ for } x \in [0,1/6],$$

and $\psi(t)$, $\phi(t)$ are Lebesgue integrable functions on $[0,1/6]$ satisfying

(i) $0 \le \psi(t) \le 2g_1'(2t)$,

(ii) $0 \le \phi(t) \le 2g_1'(2t+1/3)$,

(iii) $2g_1'(2t) - 3g_2'(3t+1/2) \le \psi(t) - \phi(t) \le -2g_1'(2t+1/3) + 3g_2'(3t)$,

for almost all $t \in [0,1/6]$.

Example 33. The functions $c_0(x)$, $c_1(x)$, a x solve $g_f(x) = g_h(x)$ for all $x \in [0,1]$. Putting $g_1(x) = g_2(x) = x$ in Theorem 48, the following solution

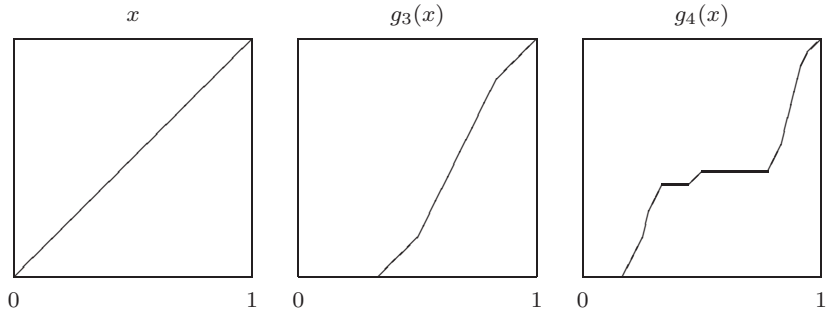
$g_3(x)$ of $g_f = g_h$ can be found:

$$g_3(x) = \begin{cases} 0 & \text{for } x \in [0, 2/6], \\ x - 1/3 & \text{for } x \in [2/6, 3/6], \\ 2x - 5/6 & \text{for } x \in [3/6, 5/6], \\ x & \text{for } x \in [5/6, 1]. \end{cases}$$

Taking $g_1(x) = g_2(x) = g_3(x)$, this $g_3(x)$ can be used as a starting point in Theorem 48 and we find

$$g_4(x) = \begin{cases} 0 & \text{for } x \in [0, 1/6], \\ 2x - 1/3 & \text{for } x \in [1/6, 3/12], \\ 4x - 5/6 & \text{for } x \in [3/12, 5/18], \\ 2x - 5/18 & \text{for } x \in [5/18, 2/6], \\ 7/18 & \text{for } x \in [2/6, 8/18], \\ x - 1/18 & \text{for } x \in [8/18, 3/6], \\ 8/18 & \text{for } x \in [3/6, 7/9], \\ 2x - 20/18 & \text{for } x \in [7/9, 5/6], \\ 4x - 50/18 & \text{for } x \in [5/6, 11/12], \\ 2x - 17/18 & \text{for } x \in [11/12, 17/18], \\ x & \text{for } x \in [17/18, 1], \end{cases}$$

see the following pictures



The study of d.f.s in $G(\{\xi(3/2)^n\})$ is also motivated by K. Mahler's (1968) conjecture: There is no $0 \neq \xi$ such that $0 \leq \{\xi(3/2)^n\} < 1/2$ for $n = 0, 1, 2, \ldots$ Mahler's conjecture follows the conjecture: Let $g(x) \in G(\{\xi(3/2)^n\})$ and $I \subset [0, 1]$. If $g(x) = \textit{constant}$ for all $x \in I$, then the length $|I| < 1/2$.

REFERENCES

[1] AMBROSIO, L.—GIGLI, N.: *A User's Guide to Optimal Transport. Modelling and Optimisation of Flows an Networks*, Lecture Notes in Mathematics. Vol. 2062, Springer-Verlag, Berlin, Heidelberg, 2013 (MR3050280).

[2] BALÁŽ, V.—MIŠÍK, L.—STRAUCH, O.—TÓTH, J. T.: *Distribution functions of ratio sequences, III*, Publ. Math. Debrecen **4770** (2013), 1-19.

[3] BALÁŽ, V.—NAGASAKA, K.—STRAUCH, O.: *Benford's law and distribution functions of sequences in* $(0, 1)$, Mathematical Notes **88** (2010), no. 4, 485–501.

[4] BARONE, H .G.: *Limit points of sequences and their transforms by methods of summability*, Duke Math. J. **5** (1939), 740–752 (Zbl. 0022.21902).

[5] BENFORD, F.: *The law of anomalous numbers*, Proc. Amer. Phil. Soc. **78** (1938), 551–572 (Zbl 18, 265; JFM 64.0555.03).

[6] BEREND, D.—BOSHERNITZAN, M. D.—KOLESNIK, G.: *Distribution modulo* 1 *of some oscillating sequences, II*, Israel J. Math. **92** (1995), 125–147.

[7] COQUET, J.—LIARDET P.: *A metric study involving independent sequences*, J. Analyse Math. **49** (1987), 15–53 (MR 89e:11043).

[8] DIACONIS, P.: *The distribution of leading digits and uniform distribution mod 1*, The Annals of Probability **5** (1977), no. 1, 72–81.

[9] DRMOTA, M.— TICHY, R. F.: *Sequences, Discrepancies and Applications*, Lecture Notes in Mathematics Vol. 1651, Springer-Verlag, Berlin, Heidelberg, 1997.

[10] FAST, H.: *Sur la convergence statistique*, Colloq. Math. **2** (1951), 241–244.

[11] FIALOVÁ, J.—STRAUCH, O.: *On two-dimensional sequences composd by one--dimensional uniformly distributed seqiuences*, Uniform Distribution Theory 6 (2011), no. 1, 101–125.

[12] FIALOVÁ, J.—MIŠÍK, L.— O. STRAUCH, O.: *An asymptotic distribution function of the three-dimensional shifted van der Corput sequence*, Applied Mathematics **5** (2014), 2334–2359, http://dx.doi.org/10.4236/am.2014.515227

[13] FILIP F.—MIŠÍK, L.—TÓTH J. T.: *On distribution functions of certain block sequences*, Uniform Distribution Theory **2** (2007), no. 1, 115–126.

[14] FRIDY, J. A.: *Statistical limit points*, Proc. Amer. Math. Soc. **118** (1993), 1187–1192.

[15] GIULIANO ANTONINI, R.—STRAUCH, O.: *On weighted distribution functions of sequences*, Uniform distribution Theory **3** (2008), no. 1, 1–18.

[16] GRABNER, P. J.—TICHY, R. F.: *Remarks on statistical independence of sequences*, Math. Slovaca **44** (1994), 91–94 (MR 95k:11098).

[17] HLAWKA, E.: *The Theory of Uniform Distribution*, A B Academic Publishers, Berkhamsted, 1984 (the translation of the original German edition Hlawka 1979) (MR 85f:11056).

[18] HOFER, M.—IACÓ, M. R.: *Optimal bounds for integrals with respect to copulas and applications*, J. Optim. Theory Appl. **161** (2014), no. 3, 999–1011 (DOI 10.1007/s10957-013-0454-x).

[19] KOKSMA, J. F.: *Asymptotische verdeling van getallen modulo 1. I, II, II,* Mathematica (Leiden) **1** (1933), 245–248; **2** (1933), 1–6, 107–114 (Zbl 0007.33901).

[20] KOKSMA, J. F.: *Diophantische Approximationen.* In: Ergebnisse der Mathematik und Ihrer Grenzgebiete, Vol. IV, Springer, Berlin, 1936 (Zbl 12, 39602).

[21] KOSTYRKO, P.—MAČAJ, M.—ŠALÁT, T.—STRAUCH, O.: *On statistical limit points*, Proc. Amer. Math. Soc. **129** (2001), no. 9, 2647–2654, (MR 2002b:40003).

[22] KUIPERS, L.—NIEDERREITER, H.: *Uniform Distribution of Sequences*, John Wiley & Sons, New York, 1974; Reprint: Dover Publications, Inc. Mineola, New York, 2006.

[23] LUCA, F.—STĂNICĂ, P.: *On the first digits of the Fibonacci numbers and their Euler function*, Uniform Distribution Theory **9** (2014), no. 1, 21–25.

[24] MYERSON, G.: *A sampler of recent developments in the distribution of sequences*, Number theory with an emphasis on the Markoff spectrum (Provo, UT, 1991), Lecture Notes in Pure and App. Math. **147**, Marcel Dekker, New York, 1993, 163–190.

[25] NIEDERREITER, H.: *Random Number Generators and Quasi–Monte Carlo Methods*. In: CBMS-NSF Regional Conference Series in Applied Mathematics **63**, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[26] NELSEN, R. B.: *An Introduction to Copulas. Properties and Applications*, Lecture Notes in Statistics **139** Springer-Verlag, New York, 1999.

[27] OHKUBO, Y.: *On sequences involving primes* Uniform Distribution Theory, **6** (2011), no. 2, 221–238.

[28] OHKUBO, Y.—STRAUCH, O.: *Distribution of leading digits of numbers*, Uniform Distribution Theory (accepted).

[29] PAVLOV, A.I.: *On distribution of fractional parts and Benford's law* Izv. Aka. Nauk SSSR Ser. Mat. **45** (1981), no. 4, 760–774. (Russian)

[30] PILLICHSHAMMER, F.—STEINERBERGER, S.: *Average distance between consecutive points of uniformly distributed sequences*, Uniform Distribution Theory **4** (2009), no. 1, 51–67.

[31] PJATECKIĬ-ŠAPIRO, I. I.: *On the laws of distribution of the fractional parts of an exponential function*, Izv. Akad. Nauk SSSR, Ser. Mat. **15** (1951), 47–52 (Russian) (MR **13**, 213d).

[32] PÓLYA, G.—SZEGŐ, G.: *Aufgaben und Lehrsätze aus der Analysis* Vol.1,2; 3rd corr. ed. Grundlehren d. math. Wiss., Band 19, Springer Verlag, Berlin, Göttingen, Heidelberg, New York, 1964.

[33] RAUZY, G.: *Propriétés statistiques de suites arithmétiques*, *Le Mathmaticien* Vol **15**, Collection SUP, Presses Universitaires de France, Paris, 1976, 133 pp. (MR **53**#13152).

[34] SCHOENBERG, I. J.: *The integrability of certain functions and related summability methods*, Amer. Math. Monthly **66** (1959), 361–375.

[35] SKLAR, M.: *Fonctions de répartition à n dimensions et leurs marges*, Publ. Inst. Statis. Univ. Paris **8** (1959), 229–231 (MR **23**# A2899).

[36] STRAUCH, O.: *Uniformly maldistributed sequences in a strict sense*, Monatsh. Math. **120** (1995), 153–164 (MR 96g:11095).

[37] STRAUCH, O.: *On set of distribution functions of a sequence*. In: *Proc. Conf. Analytic and Elementary Number Theory*, in Honor of E. Hlawka's 80th Birthday, Vienna, July 18-20, 1996, Universität Wien and Universität für Bodenkultur (W. G. Nowak and J. Schoißengeier, eds.) Vienna, 1997, 214–229 (Zbl 886.11044).

[38] STRAUCH, O.: L^2 *discrepancy*, Math. Slovaca **44** (1994), 601–632 (MR 96c:11085).

[39] STRAUCH, O.: *A new moment problem of distribution functions in the unit interval*, Math. Slovaca **44** (1994), no. 2, 171–211.

[40] STRAUCH, O.: *On distribution functions of* $\xi(3/2)^n$ mod 1, Acta Arith. **LXXXI** (1997), no. 1, 25–35 (MR 98c:11075).

[41] STRAUCH, O.: *Moment problem of the type* $\int_0^1 \int_0^1 F(x,y)\mathrm{d}g(x)\mathrm{d}g(y) = 0$. In: Proceedings of the International Conference on Algebraic Number Theory and Diophantine Analysis held in Graz, August 30 to September 5, 1998 (F. Halter-Koch, R.F. Tichy eds.), Walter de Gruyter, Berlin, New York, 2000, 423–443 (MR 2001d:11079).

[42] STRAUCH, O.: *Unsolved Problems*, in: Unsolved Problems Section on the homepage of Uniform Distribution Theory, `www.boku.ac.at/MATH/udt/unsolvedproblems.pdf`; Tatra Mt. Math. Publ. 56 (2013), 109–229 (DOI: 10.2478/tmmp-2013-0029).

[43] STRAUCH, O.—BLAŽEKOVÁ, O.: *Distribution of the sequence* p_n/n mod 1, Uniform Distribution Theory **1** (2006), 45–63.

[44] STRAUCH, O.—PORUBSKÝ, Š.: *Distribution of Sequences: A Sampler*, Peter Lang, Frankfurt am Main, 2005; Electronic revised version December 11, 2013, `https//math.boku.ac.at/udt`

[45] STRAUCH, O.—TÓTH, J. T.: *Distribution functions of ratio sequences*, Publ. Math. Debrecen **58** (2001), no. 4, 751–778.

[46] VAN DER CORPUT, J. G.: *Verteilungsfunktionen I–VIII*, Proc. Akad. Amsterdam **38** (1935), 813–821, 1058–1066, **39** (1936), 10–19, 19–26, 149–153, 339–344, 489–494, 579–590.

[47] WINTNER, A.: *On the cyclical distribution of the logarithms of the prime numbers*, Quart. J. Math. Oxford **6** (1935), no. 1, 65–68 (Zbl 11, 149).

[48] UCKELMANN, L.: *Optimal couplings between one-dimensional distributions*, Distribution with given Marginals and Moment Pproblems, (V. Beneš and J. Štěpán eds.) (Prague, 1996) 275–281. Kluwer Acad. Publ., Dordrecht, 1997, MR1614681 (99a:60012).

[49] TICHY, R. F.—THONHAUSER, S.—STRAUCH, O.—IACÓ, M. R.—BALÁŽ, V.: *Extremes of* $\int_0^1 \int_0^1 F(x,y)\mathrm{d}_x\mathrm{d}_y g(x,y)$ (submitted).

[50] WINKLER, R.: *On the distribution behaviour of sequences*, Math. Nachr. **186** (1997), 303–312.

Received February 2, 2015
Accepted November 2, 2015

Oto Strauch
Mathematical Institute
Slovak Academy of Sciences
Štefánikova 49
814 73 Bratislava
SLOVAK REPUBLIC
E-mail: oto.strauch@mat.savba.sk

Uniform Distribution Theory **10** (2015), no.2, 185–203

uniform distribution theory

ON THE DISTRIBUTION OF THE ARGUMENT OF THE RIEMANN ZETA-FUNCTION ON THE CRITICAL LINE

Selin Selen Özbek—Jörn Steuding

Dedicated to Prof. H. Niederreiter on the occasion of his 70th birthday
and
to Prof. E. Wegert on the occasion of his 60th birthday

ABSTRACT. We investigate the distribution of the argument of the Riemann zeta-function on arithmetic progressions on the critical line. We prove uniform distribution modulo $\frac{\pi}{2}$ and we show uniform distribution modulo π under certain restrictions. We also discuss continuous uniform distribution.

Communicated by Werner Georg Nowak

1. Introduction and statement of the main results

The Riemann zeta-function ζ plays a prominent role in multiplicative number theory. One of the reasons is the link between its complex zeros and the distribution of prime numbers. The famous yet unsolved Riemann hypothesis states that all so-called nontrivial (non-real) zeros $\rho = \beta + \mathrm{i}\gamma$ lie on the critical line $\frac{1}{2} + \mathrm{i}\mathbb{R}$. There are infinitely many such zeros; namely, if $N(T)$ counts there number with imaginary part $\gamma \in (0, T)$ (according multiplicities), then the Riemann-von Mangoldt formula states

$$N(T) = \frac{T}{2\pi} \log \frac{T}{2\pi e} + O(\log T), \tag{1}$$

as T tends to infinity. The main term of this formula had been announced (without proof) by Riemann in 1859 in his path-breaking nine pages article [30]; an asymptotic formula had been proved by von Mangoldt, first with an error term of size $O\big((\log T)^2\big)$ in [24], and in [25] finally with the error term given above.

2010 Mathematics Subject Classification: 11M06, 11M26.
Keywords: Riemann Zeta-Function, Uniform Distribution.

The reasoning behind these formulae is the simple principle of the argument and rather subtle properties of the zeta-function; the most important ingredient is the functional equation

$$\zeta(s) = \Delta(s)\zeta(1-s), \tag{2}$$

where

$$\Delta(s) = \pi^{s-\frac{1}{2}} \frac{\Gamma(\frac{1-s}{2})}{\Gamma(\frac{s}{2})},$$

and Γ denotes Euler's gamma-function. This symmetry was conjectured by Euler and first proved by Riemann [30]; see also Titchmarsh's monography [37]. By Stirling's formula, the asymptotic growth of Δ is for fixed real part given by

$$\Delta(\sigma + \mathrm{i}t) = \left(\frac{t}{2\pi}\right)^{\frac{1}{2}-\sigma-\mathrm{i}t} \exp\Big(\mathrm{i}\left(t + \tfrac{\pi}{4}\right)\Big)\Big(1 + O\left(t^{-1}\right)\Big), \tag{3}$$

as $t \to +\infty$. A proof of Stirling's formula can be found in Rudin's wonderful book [31], §8.22; details about its application to Δ can be found in Titchmarsh [37], §7.4. It is essentially this asymptotical formula which yields the main term in (1). Backlund [2, 3] obtained the more precise expression

$$N(T) = \frac{T}{2\pi}\log\frac{T}{2\pi e} + \frac{7}{8} + S(T) + O(T^{-1}), \tag{4}$$

where

$$S(t) := \frac{1}{\pi}\arg\zeta(\tfrac{1}{2} + \mathrm{i}t)$$

is the argument of the zeta-function on the critical line. In view of the multivalued complex logarithm one defines the value of the logarithm $\log\zeta$ at $\frac{1}{2} + \mathrm{i}t$ by continuous variation along the polygon with vertices $2, 2+\mathrm{i}t, \frac{1}{2}+\mathrm{i}t$, provided t is not equal to an ordinate of a nontrivial zero $\beta + \mathrm{i}\gamma$; otherwise, when $t = \gamma$, we define $S(\gamma)$ by

$$S(\gamma) = \tfrac{1}{2}\lim_{\epsilon\to 0}\big(S(\gamma+\epsilon) + S(\gamma-\epsilon)\big). \tag{5}$$

Notice that $\zeta(2)$ is a positive real number (equal to $\frac{\pi^2}{6}$ as Euler proved), hence $\log\zeta(s)$ is the principal branch of the logarithm on the subinterval $(1,\infty)$ of the real axis; however, for complex $\frac{1}{2} + it$ the argument might be very large. For example, Tsang [39] showed that $S(t) = o\big((\log t/\log\log t)^{\frac{1}{3}}\big)$ cannot be true.[1]

[1]Here the notation $f = o(g)$ means that the limit $\lim f(t)/g(t)$ exists as t tends to infinity and is equal to zero.

On the contrary, we have $S(t) = O(\log t)$ by von Mangoldt's work [25] and, assuming the Riemann hypothesis, the slightly better $S(t) = O(\log t/\log\log t)$ due to Littlewood [23] who also showed unconditionally that

$$\int_0^T S(t)\mathrm{d}t = O(\log T),$$

which implies that $S(t)$ oscillates heavily.

In view of (4) it follows that $S(t)$ is continuous except for ordinates of non-trivial zeros, and that $\pi S(t)$ jumps at each ordinate by an integer multiple of π (according to the multiplicity of the zero). It is conjectured that all (or at least almost all) zeta zeros are simple, so we shall expect that $\pi S(t)$ jumps by $\pm\pi$ at the ordinate of a zero (which is also reflected in Figure 2 below by the curve $t \mapsto \zeta(\frac{1}{2} + \mathrm{i}t)$ passing through the origin).

Here we are concerned about the distribution of values of the argument of the zeta-function. Although a probabilistic limit law for the logarithm of the zeta-function had already been found by Bohr[2] & Jessen [5] there are several phenomena unclarified. In view of the beautiful phase plots for complex-valued functions in general and the zeta-function in particular provided by Wegert and Semmler [32, 41] (see Figure 1) it is an interesting question whether the values of the argument are uniformly distributed in some sense. Taking the Dirichlet series representation $\zeta(s) = \sum_{n\geq 1} n^{-s}$, valid in the half-plane $\mathrm{Re}\, s > 1$, into account, $\zeta(s)$ is almost periodic in $\mathrm{Re}\, s > 1$ which is visually supported by Figure 1 and further computations (which had also been the motivation for Steuding & Wegert [36]). Notice that Bohr [4] introduced the notation of almost-periodicity in order to investigate the Riemann zeta-function and Riemann's hypothesis in particular.

In this note we shall investigate the discrete question of uniform distribution of the argument on arithmetic progressions on the critical line (although our results also provide information for the continuous case; see Corollary 4). Our main result is the following

Theorem 1. *The argument of the Riemann zeta-function on an arbitrary infinite arithmetic progression on the critical line is uniformly distributed modulo* $\frac{\pi}{2}$.

For the sake of completeness we recall the definition of uniform distribution introduced by Weyl [42] in 1916. Let μ be a positive real number. Then a sequence of real numbers x_n is said to be uniformly distributed modulo μ if for all α, β

[2]More precisely, it was Harald Bohr, the younger brother of the physicist and Nobel prize laureate Niels Bohr.

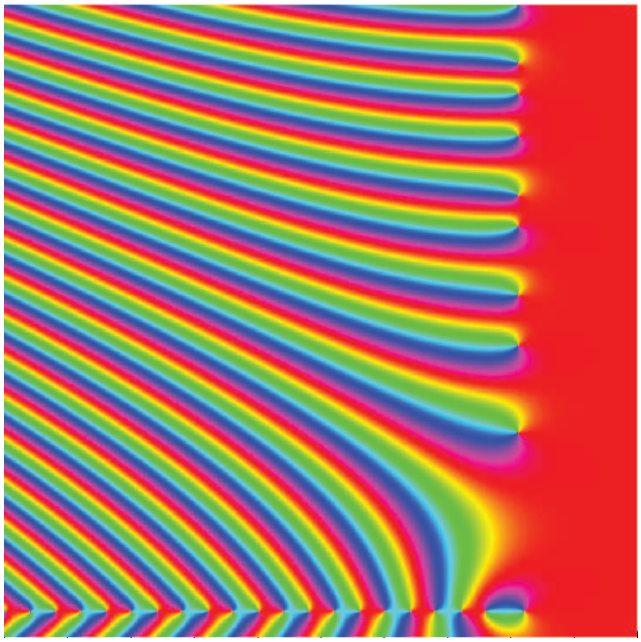


FIGURE 1. This phase plot of the zeta-function is taken from Semmler & Wegert [32] and shows the phase plot of the Riemann zeta function in the rectangle $-40 \le \operatorname{Re} s \le 10, -2 \le \operatorname{Im} s \le 48$. One can easily identify the pole of $\zeta(s)$ at $s = 1$, several nontrivial and trivial (real) zeros. One also observes a certain regularity in the colours as $\operatorname{Im} s$ increases which corresponds to almost periodicity properties of $\zeta(s)$. Some comments to this phenomenon are given in the final section.

with $0 \le \alpha < \beta \le \mu$ the proportion of the fractional parts of the x_n modulo μ in the interval $[\alpha, \beta)$ corresponds to its length in the following sense:

$$\lim_{N\to\infty} \frac{1}{N} \sharp\{1 \le n \le N \,:\, x_n \bmod \mu \in [\alpha, \beta)\} = \frac{\beta - \alpha}{\mu}.$$

Here we define $x_n \bmod \mu$ by

$$x_n \bmod \mu := x_n - \left\lfloor \frac{x_n}{\mu} \right\rfloor \mu, \tag{6}$$

where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. Weyl considered the case $\mu = 1$, however, for our purpose a modulus related to the geometry of the complex plane (i.e., a modulus μ such that $\frac{2\pi}{\mu}$ is a positive integer) is more natural.

Unfortunately, Theorem 1 does not give information about uniform distribution modulo π or 2π; for these moduli our approach only allows us to prove conditional or rather non-explicit results.

Theorem 2. *Let τ and δ be real numbers, δ positive. Assume that the number $m(N)$ of zeros of $\zeta(s)$ in the arithmetic progression $\frac{1}{2}+\mathrm{i}(\tau+n\delta)$ with $n \in \mathbb{N}$ and $n \leq N$ satisfies*

$$m(N) = o(N) \tag{7}$$

as $N \to \infty$. Then the argument of the Riemann zeta-function on the arithmetical progression $\frac{1}{2}+\mathrm{i}(\tau+n\delta)$ with $n \in \mathbb{N}$ is uniformly distributed modulo π.

It is exactly the question about nontrivial zeros of the zeta-function in arithmetic progression which makes a difference for our approach. In 1954, Putnam [27, 28] showed that there is no infinite arithmetic progression of nontrivial zeros; a sharpened result due to van Frankenhuijsen [11] states that for a hypothetical arithmetic progression with $\zeta(\frac{1}{2}+\mathrm{i}n\delta) = 0$ for $1 \leq n < N$ and fixed positive real δ its length is bounded by $N < 13\delta$. Recently, Martin & Ng [26] and Li & Radziwiłł [22] gave quantitive bounds for the hypothetical number of nontrivial zeros in an arithmetic progression, however, their results are too weak to have any impact on our reasoning. Assuming the Riemann hypothesis in addition with Montgomery's pair correlation conjecture Ford, Soundararajan & Zaharescu [10], have shown that the proportion of zeros in an arbitrary vertical arithmetic progression is indeed zero. Consequently, under this assumption (7) is fulfilled. Probably, one should not expect any zeros in vertical arithmetic progression (of length at least three). Another stronger conjecture about the zeros of $\zeta(s)$ due to Ingham [14] deals with linear independence over $\mathbb{Q}$: if the nontrivial zeros are denoted as $\beta+\mathrm{i}\gamma$, it is widely believed that there are no rational linear relations for the imaginary parts γ apart from the trivial ones (resulting from the fact that with $\beta+\mathrm{i}\gamma$ also its complex conjugate $\beta-\mathrm{i}\gamma$ is a zero). It is expected that the imaginary parts of the nontrivial zeros should not carry any arithmetical information but are distributed like random data (e.g., the eigenangles of the Gaussian Unitary Ensemble in Random Matrix Theory, see Keating Snaith [18]). In view of this difficulty with zeros in an arithmetic progression we shall also show

Theorem 3. *For all pairs $(\tau,\delta) \in \mathbb{R}\times\mathbb{R}_{>0}$ except a possible at most countable set of exceptions, the argument of the Riemann zeta-function on the arithmetical progression $\frac{1}{2}+\mathrm{i}(\tau+n\delta)$ with $n \in \mathbb{N}$ is uniformly distributed modulo π.*

The article is organized as follows. We recall some basic facts from uniform distribution theory in the following section. In Section 3 we outline our method and prove Theorem 1. The proofs of Theorem 2 and 3 are given in Section 4

as well as an application to the case of continuous distribution modulo π. In the final section we discuss related results.

2. Variations of Weyl's criterion

Weyl's celebrated criterion [42] states that a sequence of real numbers x_n is uniformly distributed modulo one if, and only if, for all integers $m \neq 0$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{1\leq n\leq N} \exp(2\pi \mathrm{i} m x_n) = 0.$$

Since we investigate the argument of a complex-valued function it is natural to have moduli related to π, and by our approach (as will be explained in the beginning of the following section) we are forced to consider $\mu = \frac{\pi}{2}$ and $\mu = \pi$, respectively. We shall make use of the following variant of Weyl's criterion: a sequence of real numbers x_n is uniformly distributed modulo μ if, and only if, for all integers $m \neq 0$,

$$\sum_{n\leq N} \exp\left(\tfrac{2\pi}{\mu} \mathrm{i} m x_n\right) = o(N), \tag{8}$$

as $N \to \infty$. For a proof of Weyl's critierion in this form we refer to Weber's monography [40], Theorem 1.6.3.

For technical reasons it is sometimes advantageous to get rid of the first elements of the sequence in question. This leads to an equivalent form of the above criterion where (8) is replaced by

$$\sum_{M<n\leq M+N} \exp\left(\tfrac{2\pi}{\mu} \mathrm{i} m x_n\right) = o(N), \tag{9}$$

which has to hold for all sufficiently large M. The changed range of summation does not influence uniform distribution since this is a property of the infinite tail of a sequence and not of finitely many initial values.

Finally, we notice that in view of the Weyl criterion in whatever form we can omit a sufficiently small set of elements from our sequence. Namely, if $\mathcal{N} \cup \mathcal{M} = \mathbb{N}$ is a disjoint union with

$$\lim_{N\to\infty} \frac{1}{N} \sharp\{n \leq N \,:\, n \in \mathcal{M}\} = 0,$$

then $(x_n)_{n\in\mathbb{N}}$ is uniformly distributed modulo μ if, and only if, the subsequence $(x_n)_{n\in\mathcal{N}}$ is uniformly distributed modulo μ. This follows immediately

from Weyl's criterion and the trivial bound

$$\sum_{\substack{n\le N\\ n\in\mathcal{M}}} \exp\left(\tfrac{2\pi}{\mu}\mathrm{i}mx_n\right) \ll \sharp\{n\le N\,:\,n\in\mathcal{M}\} \ll \epsilon N$$

for all positive ϵ and sufficiently large N.[3] Such a set $\mathcal{M}$ will be called *negligible*.

3. The main idea of our method and the proof of Theorem 1

We want to study the distribution of the argument of the zeta-function on an infinite vertical arithmetic progression:

$$\zeta\left(\tfrac{1}{2}+\mathrm{i}t_n\right) \qquad \text{with} \qquad t_n := \tau + n\delta \qquad \text{for} \quad n\in\mathbb{N}.$$

Here and in the sequel τ and δ are arbitrary but fixed real numbers, δ being positiv. In view of (9) we put $x_n = \arg\zeta\big(\frac{1}{2}+\mathrm{i}(\tau+n\delta)\big) = \pi S(\tau+n\delta)$ and thus have to study the exponential sum

$$\Sigma_m(N) \;:=\sum_{M<n\le M+N} \exp\Big(\tfrac{2\pi}{\mu}\mathrm{i}m\arg\zeta\big(\tfrac{1}{2}+\mathrm{i}(\tau+n\delta)\big)\Big), \tag{10}$$

where m is a non-zero integer and μ equals either π or $\frac{\pi}{2}$.

Firstly, we may assume that $\zeta(\frac{1}{2}+\mathrm{i}t_n)\neq 0$. By the reflection principle, $\overline{\zeta(s)}=\zeta(\overline{s})$, hence also $\zeta(\frac{1}{2}-\mathrm{i}t_n)$ does not vanish. This implies

$$\exp\big(2m\mathrm{i}\arg\zeta(\tfrac{1}{2}+\mathrm{i}t_n)\big) = \left(\frac{\zeta(\frac{1}{2}+\mathrm{i}t_n)}{|\zeta(\frac{1}{2}+\mathrm{i}t_n)|}\right)^{2m} = \left(\frac{\zeta(\frac{1}{2}+\mathrm{i}t_n)}{\zeta(\frac{1}{2}-\mathrm{i}t_n)}\right)^{m}.$$

Notice that the quantity on the right-hand side determines the argument of the zeta-function only modulo 2π (the period of the exponential function $t\mapsto\exp(\mathrm{i}t)$). Consequently, our reasoning is limited to positive real moduli μ for which $\frac{2\pi}{\mu}$ is an integer. This gives for the corresponding exponential sum in Weyl's criterion (9)

$$\Sigma_m(N) \;=\sum_{M<n\le M+N} \left(\frac{\zeta(\frac{1}{2}+\mathrm{i}t_n)}{\zeta(\frac{1}{2}-\mathrm{i}t_n)}\right)^{\frac{\pi}{\mu}m}. \tag{11}$$

Now, using the functional equation (2), we get

$$\Sigma_m(N) \;=\sum_{M<n\le M+N} \Delta\left(\tfrac{1}{2}+\mathrm{i}t_n\right)^{\frac{\pi}{\mu}m}, \tag{12}$$

[3]Here we have used Vinogradov-notation $f\ll g$ which is equivalent to $f=O(g)$.

which is not too difficult to estimate (see (14) and its proof below). It will become clear soon that by this method the modulus $\mu = 2\pi$ is impossible to combine with odd m.

In view of the above simplification (12) one might be tempted to consider our results as easy consequences of the functional equation only. However, there is a certain obstacle, namely our assumption on the non-vanishing of the zeta-function on the arithmetic progression in question.

Now let $\mu = \frac{\pi}{2}$. If $\zeta(\frac{1}{2} + \mathrm{i}t_n) = 0$, i.e., $t_n = \gamma$ is an ordinate of a nontrivial zero, then the argument of the zeta-function is in view of (5) given by

$$\pi S(\gamma) = \tfrac{\pi}{2} \lim_{\epsilon \to 0} \big(S(\gamma + \epsilon) + S(\gamma - \epsilon) \big)$$

(and it is this definition which appears the reason for our restriction to $\mu = \frac{\pi}{2}$ in Theorem 1). In this case the values $S(\gamma \pm \epsilon)$ differ by an integer (as follows from (4) and had already been explained in the introductory section). Thus, for every sufficiently small ϵ,

$$\lim_{\epsilon \to 0} \pi S(\gamma + \epsilon) \equiv \lim_{\epsilon \to 0} \pi S(\gamma - \epsilon) \bmod \pi,$$

respectively,

$$\pi S(\gamma) = \tfrac{1}{2} \lim_{\epsilon \to 0} \big(2\pi S(\gamma - \epsilon) + \pi k \big)$$

for some integer k. Hence, we obtain

$$\begin{aligned} \exp\Big(4m\mathrm{i} \arg \zeta \big(\tfrac{1}{2} + \mathrm{i}\gamma \big) \Big) &= \exp\Big(4m\mathrm{i} \lim_{\epsilon \to 0} \pi S(\gamma - \epsilon) \Big) \exp\big(4m\mathrm{i} \tfrac{\pi k}{2} \big) \\ &= \lim_{\epsilon \to 0} \exp\Big(4m\mathrm{i} \arg \zeta \big(\tfrac{1}{2} + \mathrm{i}(\gamma - \epsilon) \big) \Big) \\ &= \lim_{\epsilon \to 0} \Delta\big(\tfrac{1}{2} + \mathrm{i}(\gamma - \epsilon) \big)^{2m} \end{aligned}$$

by the reasoning from above. Since the limit on the right-hand side exists, we get

$$\exp\Big(4m\mathrm{i} \arg \zeta \big(\tfrac{1}{2} + \mathrm{i}\gamma \big) \Big) = \Delta\big(\tfrac{1}{2} + \mathrm{i}\gamma \big)^{2m}.$$

Consequently, for the modulus $\mu = \frac{\pi}{2}$ it does not matter whether $\frac{1}{2} + \mathrm{i}t_n$ is a zero of the zeta-function or not (since $\pi S(t)$ is continuous modulo $\frac{\pi}{2}$). Notice that the above reasoning fails for $\mu = \pi$ because in this case our reasoning would lead to $\exp\big(2m\mathrm{i} \arg \zeta(\frac{1}{2} + \mathrm{i}\gamma)\big) = \pm\Delta(\frac{1}{2} + \mathrm{i}\gamma)^m$ without control of the sign $\pm$.

In view of (12) we arrive at

$$\Sigma_m(N) = \sum_{M < n \le M+N} \Delta\big(\tfrac{1}{2} + \mathrm{i}t_n \big)^{2m}.$$

Taking (3) into account this expression can be simplified to

$$\Sigma_m(N) = \Big(\exp\big(\tfrac{\pi m}{2}\mathrm{i}\big) + O\big(M^{-|2mk|}\big)\Big) \sum_{M<n\le M+N} \left(\frac{2\pi e}{t_n}\right)^{2mit_n}. \tag{13}$$

Next we apply van der Corput's method [7] for estimating the appearing exponential sum. For this aim we shall make use of Lemma 8.12 from Iwaniec & Kowealski's monography [15]: let $b-a \ge 1$, let f be a twice differentiable real-valued function on $[a,b]$ with $f''(x) \ge \Lambda > 0$ on $[a,b]$. Then

$$\sum_{a<n<b} \exp\big(2\pi\mathrm{i} f(n)\big) \ll \big(f'(b) - f'(a) + 1\big)\Lambda^{-\frac{1}{2}},$$

where the implicit constant is absolute. Of course, the statement of this lemma is also true if $f''(x) \le -\Lambda < 0$ on $[a,b]$ (as follows by complex conjugation); in this case we only have to change the upper bound to

$$\big(f'(a) - f'(b) + 1\big)\Lambda^{-\frac{1}{2}}.$$

In view of (13) we consider

$$f(x) = -\frac{m}{\pi}(\tau + x\delta)\log\frac{\tau + x\delta}{2\pi e}$$

with

$$f'(x) = -\frac{m\delta}{\pi}\log\frac{\tau + x\delta}{2\pi} \qquad \text{and} \qquad f''(x) = -\frac{m\delta^2}{\pi(\tau + x\delta)}.$$

Applying the lemma with $a = M$, $b = M + N$ and Λ of approximate size N^{-1} leads to

$$\Sigma_m(N) \ll N^{\frac{1}{2}}\log N, \tag{14}$$

where the implicit constant may depend on τ, δ and m. In view of (9) this estimate is suitable and the theorem is proved. Notice that we do not attempt to prove sharper bounds here.

4. Proof of Theorem 2 and 3 and an application to continuous uniform distribution

First we prove Theorem 2. In view of (7) the set of elements $\frac{1}{2} + \mathrm{i}t_n$ of our arithmetic progression which leads to zeros of the zeta-function is (according to our remark in Section 2) negligible with respect to uniform distribution modulo π. Hence, we may assume that $\zeta(\frac{1}{2} + \mathrm{i}t_n)$ does not vanish for any positive integer n. Since $\arg\zeta(\frac{1}{2} + \mathrm{i}t) = \pi S(t)$ makes jumps at the ordinates which are integer

multiples of π, we obtain by the same reasoning as in the previous proof applied to $f(x) = -\frac{m}{2\pi}(\tau + x\delta)\log\frac{\tau+x\delta}{2\pi e}$ uniform distribution modulo π.

The proof of Theorem 3 is just an application of Cantor's notions of countability and uncountability. Since the set of nontrivial zeros is countable and the set of $(\tau, \delta) \in \mathbb{R} \times \mathbb{R}_{>0}$ is not, for almost all pairs of τ and δ (in the sense of Lebesgue measure) the corrseponding arithmetic progression $\frac{1}{2} + \mathrm{i}(\tau + n\delta)$ will avoid nontrivial zeros. Consequently, the statement follows by the same reasoning as in the previous proofs. Of course, the statement of the theorem also holds with either fixed $\delta = 1$ or fixed $\tau = 0$.

In view of the just proven theorem one can hardly expect that the continuous version of uniform distribution differs from the discrete one. To be more precise, since the argument $\pi S(t_n)$ of the zeta-function on almost all vertical arithmetic progressions t_n on the critical line is uniformly distributed modulo π, the values $\pi S(t)$ modulo π have to be equidistributed as well as t ranges continuously through $\mathbb{R}$. In order to prove that we start with a standard notion. A real-valued Lebesgue integrable function f defined on $(0, \infty)$ is said to be continuously uniformly distributed modulo μ if for all $0 \le \alpha < \beta \le \mu$ the following limit exists and satisfies

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{1}_{[\alpha,\beta)}(\{f(t)\})\mathrm{d}t = \frac{\beta-\alpha}{\mu},$$

where $\mathbf{1}_{[\alpha,\beta)}(f)$ is the characteristic function of the interval $[\alpha, \beta)$ (being equal to 1 if $f \in [\alpha, \beta)$ and zero otherwise) and $\{f\}$ denotes the fractional part modulo μ defined analogously to (6). This is a straightforward generalization of the corresponding definition of continuous uniform distribution modulo one which can be found, for instance, in the monograph of Kuipers & Niederreiter [19]. Moreover, their Theorem 9.6 states that a real-valued Lebesgue integrable function f is continuously uniformly distributed modulo one if the sequence of real numbers $x_n = f(\tau+n)$ with $n = 1, 2, \ldots$ is uniformly distributed modulo one for almost all τ; the proof follows from the continuous version of Weyl's criterion and incorporating the information about the discrete uniform distribution in form of approximating exponential sums. We leave it to the reader to extend this result to the situation of (continuous) uniform distribution modulo μ and just mention that in combination with Theorem 3 this reasoning leads to

Corollary 4. *The values of the argument of the zeta-function on the critical line are continuously uniformly distributed modulo π.*

A straightforward proof of this result (without the detour via discrete uniform distribution) would follow from the continuous Weyl criterion (which can be found in Kuipers & Niederreiter [19] as well as already in Weyl's original paper [42]).

5. Concluding remarks

Shanks [33] conjectured that the curve $\mathcal{C} : t \mapsto \zeta(\frac{1}{2}+it)$ approaches the origin most of the times from the third or fourth quadrant, i.e., $\zeta'(\frac{1}{2}+\mathrm{i}\gamma)$ is positive real in the mean. This was proved by Fujii [12] and Trudgian [38]. In this sense, Figure 2 (as well as Figure 3 below) provides a good idea about the distribution of the values of the zeta-function on the critical line. In the papers of Kalpokas & Steuding [16] and Kalpokas, Korolev & Steuding [17] the value-distribution of the

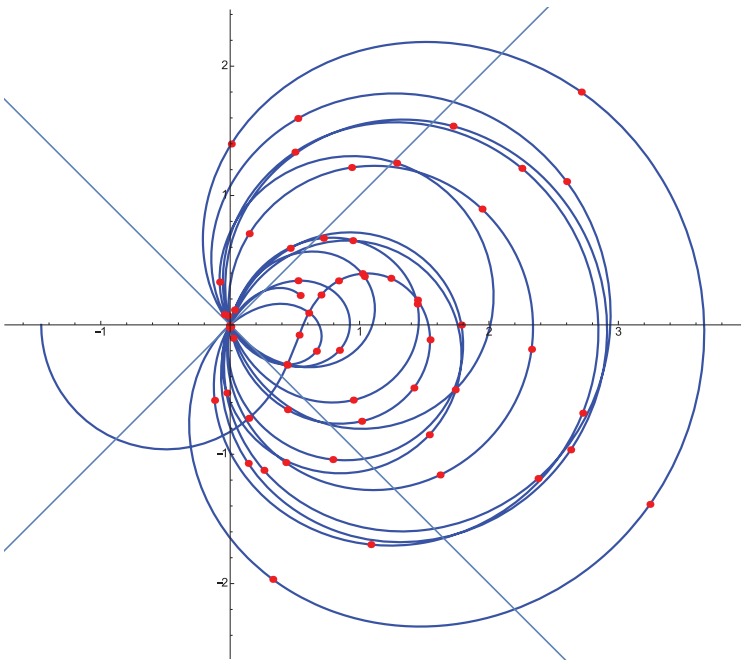


FIGURE 2. The figure shows the graph of the curve $t \mapsto \zeta\big(\frac{1}{2}+\mathrm{i}t\big)$ in the complex plane for the range $0 \le t \le \mathbf{60}$ with 13 simple zeros. The values of $\zeta(s)$ on the arithmetic progression $s = \frac{1}{2}+in$ for $n = 1, 2, \ldots, 60$ are marked as red points; the values for $n = 14, 21, 25, 33, 41, 48$ and 53 are very close to the origin (and the corresponding $\frac{1}{2}+in$ near to nontrivial zeros). In the eight octants we count in sequence $18, 7, 3, 0, 1, 3, 10, 18$ starting with the sector in the first quadrant bounded by the positive real axis and continuing counterclockwise. The data matches the statements of the theorems. The graphic resembles a similar one due to Shanks [33]. We do not provide any data about these computations here since nowadays – different to the time of Shanks, Haselgrove and others—it is easy to reproduce the data by modern computer algebra packages.

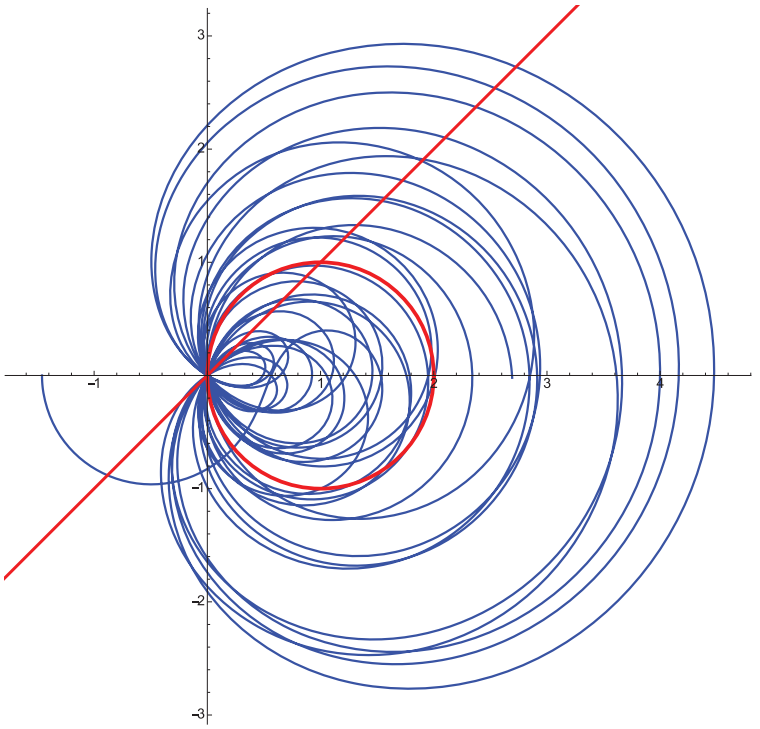


FIGURE 3. The curve $t \mapsto \zeta\left(\frac{1}{2}+\mathrm{i}t\right)$ for $t \in [0, 100]$ in addition with the circle of the mean-values and the bisecting line of the first and third quadrant ($\phi = \frac{\pi}{4}$); here the mean-value is $\frac{1}{2}(1+\mathrm{i})$ which is the intersection point (different from the origin) of the circle with the bisecting line. A different approach is due to Arias de Reyna, Brent & van der Lune [1].

curve $\mathcal{C}$ was investigated in detail. Their results explain why "*the real part of* ζ *has a strong tendency to be positive*" as observed by, for example, Edwards in his monography [8] (page 121), as well as the almost symmetry of $\mathcal{C}$ with respect to the real axis. In particular, they have shown that the mean value of the intersection points of the curve $t \mapsto \zeta(\frac{1}{2}+\mathrm{i}t)$ with an arbitrary straight line $\exp(\mathrm{i}\varphi)$ with $\phi \in [0, \pi)$ exists and is equal to $2\exp(\mathrm{i}\phi)\cos\phi = 1 + \exp(\mathrm{i}\phi)$ (see Figure 3).[4]

[4]Moreover, they showed that the zeta-function on the critical line assumes arbitrarily large positive real values and arbitrarily large negative values (what figures like Figure 2, 4 and 5 cannot show).

Notice that the circle built from these mean values, i.e., $1+\exp(\mathrm{i}\phi)$ for $0 \le \phi < \pi$, reflects the typical shape of the curve made from the zeta-values $\zeta(\frac{1}{2}+\mathrm{i}t)$ as t ranges through some fixed interval, and taking this circle as a model for the value-distribution of the zeta-function, we observe uniform distribution modulo π for a generic arithmetic progression.[5]

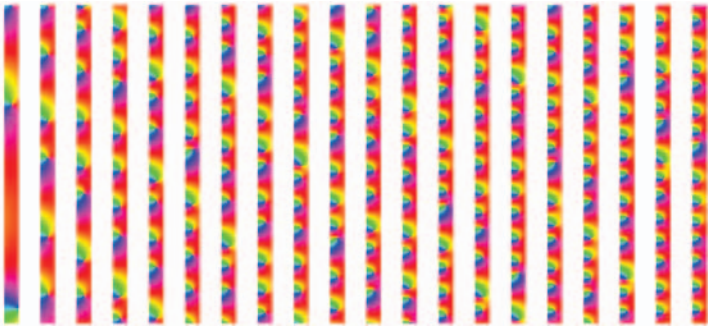


FIGURE 4. These are phase plots from [36] showing $\zeta(s)$ along the critical strip.

In Steuding & Wegert [36] it is written that "*visual inspection of the phase plot of ζ in the critical strip almost immediately reveals a surprising 'stochastic periodicity' of the phase $\psi = \zeta/|\zeta|$. (...) The eye-catching yellow diagonal stripes led us to the conjecture that the phase plot of zeta in the critical strip has sort of stochastic period, and some heuristic arguments suggest that its length equals $2\pi/\log 2$.*" According to this observation the authors proved for fixed $s \in \mathbb{C}\setminus\{1\}$ with $\operatorname{Re} s \in (0,1]$, and $d = \frac{2\pi}{\log \ell}$ with $2 \le \ell \in \mathbb{N}$,

$$\frac{1}{N}\sum_{0\le n<N} \zeta(s+\mathrm{i}nd) = (1-\ell^{-s})^{-1} + O\left(N^{-\operatorname{Re} s}\log N\right), \quad \text{as} N \to \infty.$$

It is worth to notice that the main term equals the factor for a prime number ℓ in the Euler product representation of the zeta-function $\zeta(s) = \prod_\ell (1-\ell^{-s})^{-1}$, valid for $\operatorname{Re} s > 1$. This factor on its own obeys a uniform distribution law modulo 2π with any arithmetic progression $t_n = \tau + \delta n$ for which $\delta \log \ell \notin 2\pi\mathbb{Q}$. Nevertheless, for a generic progression $t_n = \tau + n\frac{2\pi}{\log \ell}$ we have uniform distribution modulo π by Theorem 3 (and Figure 5 provides and illustration).

[5]Something similar can be deduced for the characteristic polynomials to random matrices from certain unitary matrix ensembles which are often used to model the value-distribution of $\zeta(\frac{1}{2}+\mathrm{i}t)$, see Keating & Snaith [18].

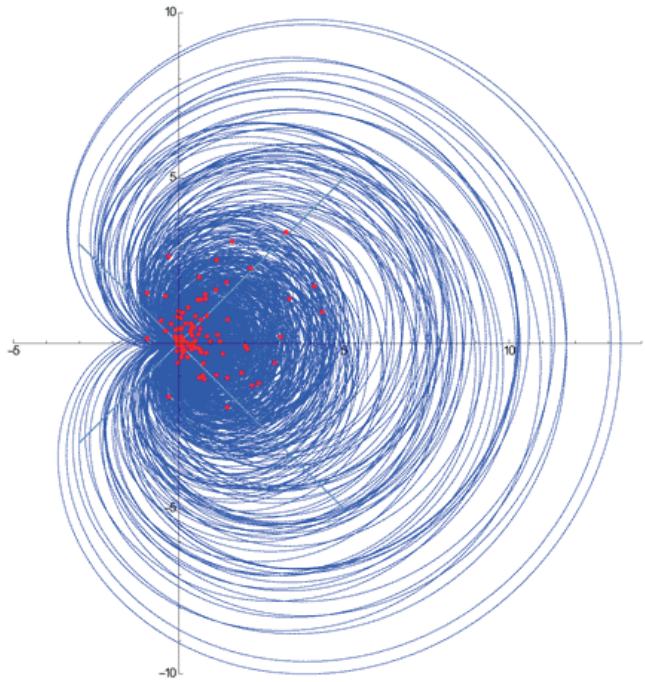


FIGURE 5. The figure shows the graph of the curve $t \mapsto \zeta\big(\frac{1}{2} + \mathrm{i}t\big)$ in the complex plane for the range $1955 \le t \le 2865$. The values of $\zeta(s)$ on the arithmetic progression $s = \frac{1}{2} + \mathrm{i}\big(1955 + n\frac{2\pi}{\log 2}\big)$ for $n = 1, 2, \dots, 100$ are marked as red points. Notice that the range up to 2865 is according to the one hundred points: $1955 + 100 \cdot \frac{2\pi}{\log 2} < 2865 < 1955 + 101 \cdot \frac{2\pi}{\log 2}$. The distributions in the four quadrants is $44, 15, 6, 35$, matching the uniform distribution modulo π very well; in the octants, however, we count $18, 26, 12, 3, 1, 5, 12, 23$ (in the same sequence as before) which shows some deviation. It is an interesting phenomenon that the values taken on this arithmetic progression are *comparably small*.

In view of the graphics and mean-value theorems mentioned above (as, e.g., in [16, 36]) one might be tempted to conjecture that the values of the argument of the zeta-function on an arithmetic progression on the critical line are not uniformly distributed modulo 2π. Actually, we are uncertain about the critical line but expect this to hold for vertical lines to the right. It is easy to show that

$$\operatorname{Re}\zeta(\sigma + \mathrm{i}t) > 2 - \zeta(\sigma) \quad \text{for} \quad \sigma > 1$$

and

$$2 > \zeta(\sigma) \quad \text{for} \quad \sigma > 1.72865.$$

Hence, the argument of $\zeta(s)$ for fixed $\operatorname{Re} s$ and varying $\operatorname{Im} s$ is definitely not uniformly distributed modulo 2π. It might be interesting to notice that, building on computer experiments, Arias de Reyna, Brent & van der Lune [1] write that it is plausible that

$$\lim_{T\to\infty} \frac{1}{T}\lambda\Big(\{t \in [0,T) \,:\, \operatorname{Re}\zeta\left(\tfrac{1}{2}+\mathrm{i}t\right) < 0\}\Big) = \frac{1}{2},$$

where λ denotes the Lebesgue measure. Unfortunately, their reasoning is limited to vertical lines to the right of the critical line. (See Figure 6 below for an illustration of this question with respect to an arithmetic progression.)

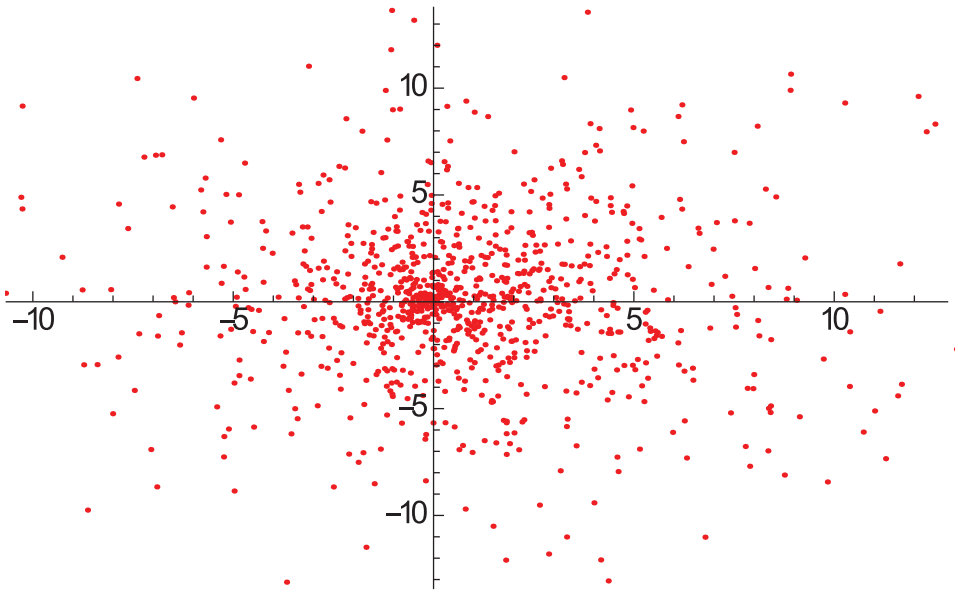


FIGURE 6. The points $\zeta(\frac{1}{2} + \mathrm{i}n\frac{2\pi}{\log 2})$ for $n = 10^{15}+1,\ldots,10^{15}+1\,000$ (without the curve). By this sample one could imagine uniform distribution modulo 2π.

We conclude with a historical detail. The first to investigate uniform distribution in the context of the zeta-function was Rademacher [29] in 1956 who proved that the ordinates of the nontrivial zeros of the zeta-function are uniformly distributed modulo one provided that the Riemann hypothesis is true;

later Elliott [9] remarked that the latter condition can be removed, and (independently) Hlawka [13] obtained the following unconditional theorem: for any real number $\alpha \neq 0$ the sequence $\alpha\gamma$, where γ ranges through the set of positive ordinates of the nontrivial zeros of $\zeta(s)$ in ascending order, is uniformly distributed modulo one. In particular, the ordinates of the nontrivial zeros of the zeta-function are uniformly distributed modulo one. In [34] the second author proved that the same holds true for the ordinates of the roots of the equation

$$\zeta(s) = a,$$

where a is an arbitrary fixed complex number. These so-called a-points had been under consideration ever since Landau's work [20, 6].[6] As a matter of fact, one distinguishes between trivial and nontrivial a-points, and indeed the trivial ones are similarly distributed as the trivial zeros of $\zeta(s)$ as $s = -2n$, $n \in \mathbb{N}$. Also the distribution of the nontrivial a-points shares some patterns with the nontrivial zeros. If $N_a(T)$ denotes the number of nontrivial a-points with imaginary part γ_a satisfying $0 < \gamma_a \leq T$ (counted according multiplicities), then pretty similar to (1) one has

$$N_a(T) = \frac{T}{2\pi} \log \frac{T}{2\pi e c_a} + O(\log T), \quad \text{as } T \to \infty,$$

where

$$c_a = 1 \quad \text{if } a \neq 1, \quad \text{and} \quad c_1 = 2.$$

One observes that the main term is independent of a (which is not too surprising having Nevanlinna's value distribution theory in mind). Moreover, Landau proved that almost all a-points are clustered around the critical line provided the Riemann hypothesis is true, and later Levinson [21] showed that this holds unconditionally. More details can be found in Steuding [35].

Acknowledgements. The authors would like to thank the anonymous referee for his or her valuable corrections and comments.

REFERENCES

[1] ARIAS DE REYNA, J.—BRENT, R. P.—VAN DE LUNE, J.: *On the sign of the real part of the Riemann zeta function*, in: *Number Theory and Related Fields. In memory of Alf van der Poorten*, (J.M. Borwein et al. eds.), Springer, New York, 2013, pp. 75–97.

[2] BACKLUND, R. J.: *Sur les zéros de la fonction $\zeta(s)$ de Riemann*, Comptes Rendus Acad. Sci. Paris **158** (1914), 1979–1981.

[6] The paper [6] of Bohr, Landau & Littlewood consists of three independent chapters, the first one due to Bohr, the second written by Landau, and the third by Littlewood.

[3] BACKLUND, R. J.: *Über die Nullstellen der Riemannschen ζ-Funktion*, Dissertation Helsingfors. Acta Math. **41**, (1916) no. 1, 345—375.

[4] BOHR, H.: *Über eine quasi-periodische Eigenschaft Dirichletscher Reihen mit Anwendung auf die Dirichletschen L-Funktionen,* Math. Ann. **85** (1922), 115–122.

[5] BOHR, H.—JESSEN, B.: , *Über die Werteverteilung der Riemannschen Zeta-funktion. II*, Acta Math. **58** (1932), 1–55.

[6] BOHR, H.—LANDAU, E.—LITTLEWOOD, J. E.: *Sur la fonction $\zeta(s)$ dans le voisinage de la droite $\sigma = \frac{1}{2}$*, Bull. de l'Acad. Royale de Belgique (1913), 3–35.

[7] VAN DER CORPUT, J. G.: *Zahlentheoretische Abschätzungen*, Math. Ann. **84** (1921), 53–79.

[8] EDWARDS, H. M.: *Riemann's Zeta function*, Pure and Applied Mathematics, Vol. 58, Academic Press, New York-London, 1974.

[9] ELLIOTT, P. D. T. A.: *The Riemann zeta function and coin tossing*, J. Reine Angew. Math. **254** (1972), 100–109.

[10] FORD, K.—SOUNDARARAJAN, K.—ZAHARESCU, A.: , *On the distribution of imaginary parts of zeros of the Riemann zeta function, II*, Math. Ann. **343** (2009), 487–505.

[11] VAN FRANKENHUIJSEN, M.: , *Arithmetic progressions of zeros of the Riemann zeta-function*, J. Number Theor. **115** (2005), 360–370

[12] FUJII, A.: *On a conjecture of Shanks*, Proc. Japan Acad. Ser. A Math. Sci. **70** (1994), 109–114

[13] HLAWKA, E.: *Über die Gleichverteilung gewisser Folgen, welche mit den Nullstellen der Zetafunktion zusammenhängen*, Österr. Akad. Wiss., Math.-Naturw. Kl. Abt. II **184** (1975), 459–471

[14] INGHAM, A. E.: *On two conjectures in the theory of numbers*, Amer. J. Math. **64** (1942), 313–319.

[15] IWANIEC, H.—KOWALSKI, E.: *Analytic Number Theory*, AMS, Providence, 2004.

[16] KALPOKAS, J.—STEUDING, J.: *On the value-distribution of the Riemann zeta-function on the critical line*, Moscow J. Combinatorics Number Theory **1** (2011), 26–42.

[17] KALPOKAS, J.—KOROLEV, M.—STEUDING, J.: *Negative values of the Riemann zeta function on the critical line*, Mathematika **59** (2013), 443–462.

[18] KEATING, J. P.—SNAITH, N. C.: *Random matrix theory and $\zeta(1/2 + it)$*, Commun. Math. Phys. **214** (2000), 57–89.

[19] KUIPERS, L.—NIEDERREITER, H.: *Uniform distribution of sequences*, John Wiley & Sons, New York, 1974.

[20] LANDAU, E.: *Über die Nullstellen der Zetafunktion*, Math. Ann. **71** (1912), 548–564.

[21] LEVINSON, N.: *Almost all roots of $\zeta(s) = a$ are arbitrarily close to $\sigma = 1/2$*, Proc. Nat. Acad. Sci. U.S.A. **72** (1975), 1322–1324.

[22] LI, X.—RADZIWIŁŁ, M.: *The Riemann zeta function on vertical arithmetic progressions*, Int. Math. Res. Not. **2** (2015), 325–354.

[23] LITTLEWOOD, J. E.: *On the zeros of the Riemann zeta-function*, Proc. Cambridge Philos. Soc. **22** (1924), 295–318.

[24] VON MANGOLDT, H.: *Zu Riemann's Abhandlung "Ueber die Anzahl der Primzahlen unter einer gegebenen Größe"*, J. Reine Angew. Math. **114** (1895), 255–305.

[25] VON MANGOLDT, H.: *Zur Verteilung der Nullstellen der Riemannschen Funktion* $\xi(t)$, Math. Ann. **60** (1905), 1–19.

[26] MARTIN, G—NG, N.: *Nonzero values of Dirichlet L-functions in vertical arithmetic progressions*, Int. J. Number Theory **9** (2013), 813–843.

[27] PUTNAM, C. R.: *On the non-periodicity of the zeros of the Riemann zeta-function*, Amer. J. Math. **76** (1954), 97–99.

[28] PUTNAM, C. R.: *Remarks on periodic sequences and the Riemann zeta-function*, Amer. J. Math. **76** (1954), 828–830.

[29] RADEMACHER, H. A.: *Fourier Analysis in Number Theory*, in: Collected Papers of Hans Rademacher, Vol. II, Symposium on Harmonic Analysis and Related Integral Transforms (Cornell Univ., Ithaca, N.Y., 1956) Massachusetts Inst. Tech., Cambridge, Mass., 1974, pp. 434–458,.

[30] RIEMANN, R.: *Über die Anzahl der Primzahlen Unterhalb Einer Gegebenen Grösse*, Monatsber. Preuss. Akad. Wiss. Berlin (1859), 671–680.

[31] RUDIN, W.: *Principles of Mathematical Analysis*, the 3rd ed. MacGraw-Hill, 1976.

[32] SEMMLER, G.—WEGERT, E.: *Phase plots of complex functions: a journey in illustration*, Notices Amer. Math. Soc. **58** (2011), no. 6, 768–780.

[33] SHANKS, D.: *Review of 'Tables of the Riemann zeta function' by C. B. Haselgrove in collaboration with J. C. P. Miller*, Math. Comp. **15** (1961), 84–86.

[34] STEUDING, J.: *The roots of the equation* $\zeta(s) = a$ *are uniformly distributed modulo one*, in: Analytic and Probabilistic Methods in Number Theory, (A. Laurinčikas et al. eds.), Proceedings of the Fifth International Conference in Honour of J. Kubilius, Palanga 2011, TEV, Vilnius 2012, 243–249.

[35] STEUDING, J.: *One hundred years uniform distribution modulo one and recent applications to Riemann's zeta-function*, in: Topics in Mathematical Analysis and Applications (T.M. Rassias, L. Tóth eds), Springer Optim. App. **94**, springer Cham (2014), pp.659–698.

[36] STEUDING, J.—WEGERT, E.: *The Riemann zeta function on arithmetic progressions*, Exp. Math. **21** (2012), 235–240.

[37] TITCHMARSH, E. C.: *The theory of the Riemann zeta-function*, the 2nd ed. (D. R. Heath-Brown, ed.), The Clarendon Press, Oxford University Press, New York, 1986.

[38] TRUDGIAN, T. S.: *On a conjecture of Shanks*, J. Number Theory **130** (2010), 2635–2638.

[39] TSANG, K.-M.: , Some Ω-theorems for the Riemann zeta-function, Acta Arith. **46** (1985), 369–395.

[40] WEBER, M.: *Dynamical Systems and Processes*, 14. European Mathematical Society (EMS), Zürich, 2009.

[41] WEGERT, E.: *Visual Complex Functions*, Birkhäuser Basel, 2012.

[42] WEYL, H.: *Über die Gleichverteilung von Zahlen Mod. Eins*, Math. Ann. **77** (1916), 313–352.

Received April 7, 2015
Accepted June 18, 2015

Selin Selen Özbek
Akdeniz University
Department of Mathematics
07058 Antalya
TURKEY
E-mail: s.selenozbek@gmail.com

Jörn Steuding
Department of Mathematics
Würzburg University
Emil-Fischer-Str. 40
97 074 Würzburg
GERMANY
E-mail: steuding@mathematik.uni-wuerzburg.de

Nair, Radhakrishnan (Liverpool) nair@liverpool.ac.uk [ergodic theoretic aspects of uniform distribution, issues related to pointwise convergence, metrical theory of uniform distribution, exceptional sets, exponential sums, distribution of primes, densities and combinatorial number theory];
Niederreiter, Harald (Linz, Salzburg) harald.niederreiter@oeaw.ac.at, ghnied@gmail.com [all parts of uniform distribution theory];
Ohkubo, Yukio (Kagoshima) ohkubo@eco.iuk.ac.jp [distribution of one dimensional and multidimensional sequences, continuous uniform distribution, discrepancies];
Paštéka, Milan (Bratislava) pasteka@mat.savba.sk [theory of densities, uniform distribution in groups and rings];
Pethő, Attila (Debrecen) petho.attila@inf.unideb.hu [diophantine equations, distribution of polynomials and algebraic numbers, radix representations, cryptography];
Pillichshammer, Friedrich (Linz) friedrich.pillichshammer@jku.at [discrepancy, digital nets, quasi-Monte Carlo integration, lattice rules, tractability of high-dimensional problems];
Porubský, Štefan (Prague) sporubsky@hotmail.com [distribution functions of sequences, combinatorial number theory, arithmetic densities, arithmetic functions, summation methods, pseudorandom number generators, cryptography];
Sárközy, András (Budapest) sarkozy@cs.elte.hu [distribution of binary sequences, pseudorandom number generators, combinatorial number theory, character sums, number theoretic ciphers];
Shkredov, Ilya (Moscow) ishkredov@rambler.ru [combinatorial number theory, continued fractions];
Sós, Vera T. (Budapest) sos@renyi.hu [uniform distribution of sequences and discrepancies];
Strauch, Oto (Bratislava) strauch@mat.savba.sk [theory of distribution functions of sequences, discrepancies, metric theory of diophantine approximations];
Tezuka, Shu (Kyushu) tezuka@math.kyushu-u.ac.jp [quasi-Monte Carlo integration, quasi-Monte Carlo methods in financial mathematics, pseudorandom number generators];
Tichy, Robert F. (Graz) tichy@tugraz.at [discrepancies, quasi-Monte Carlo integration, quasi-Monte Carlo methods in financial mathematics; diophantine approximations and equations];
Tóth, János T. (Komárno) tothj@selyeuni.sk [distribution of block sequences, various kinds of dense sequences];
Ustinov, Alexey V. (Khabarovsk) ustinov@iam.khv.ru [trigonometric (exponential) sums, continued fractions geometry of numbers];
Weber, Michel (Strasbourg) michel.weber@math.unistra.fr [ergodic theory, pointwise convergence, probability theory];
Winkler, Reinhard (Vienna) reinhard.winkler@tuwien.ac.at [distribution of integer sequences and sequences from groups and generalized spaces, the theory of distribution functions of sequences (limit measures), distribution of binary sequences, dynamics emerging from sequences];
Winterhof, Arne (Linz) arne.winterhhof@oeaw.ac.at [pseudorandom numbers, measures of pseudorandomness, exponential sums, finite fields];
Woźniakowski, Henryk (New York, Warsaw) henryk@cs.columbia.edu [discrepancies, quasi-Monte Carlo algorithms, Monte-Carlo algorithms, tractability of multivariate problems, computational complexity of continuous problems].